PCLmed: Champion Solution for ImageCLEFmedical 2024 Caption Prediction Challenge via Medical Vision-Language Foundation Models

Bang Yang[1,2], Yue Yu[1], Yuexian Zou[1,2] and Tong Zhang[1,*]

[1]Peng Cheng Laboratory, Shenzhen 518055, China
[2]ADSPLAB, School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China

Abstract

Automatically generating captions and reports for medical images has become increasingly important due to the growing workload of radiologists in hospitals. To tackle this challenging task with limited annotation data, there is a rising interest in developing medical vision-language foundation models (Med-VLFMs). These models leverage the capabilities of vision foundation models and large language models (LLMs) and often utilize the parameter-efficient fine-tuning (PEFT) technique. However, current Med-VLFMs face two critical issues: (1) relying on a single vision model to represent the semantics of medical images, and (2) adapting LLMs with PEFT without considering the interference between vision and text modalities. This work presents a novel Med-VLFM with vision encoder ensembling (VEE) and modality-aware adaptation (MAA) to address these limitations. VEE combines the strengths of general and medical specialist vision foundation models to produce a more holistic representation of medical images. MAA introduces two small sets of trainable parameters into LLMs to calibrate vision and text features, respectively. Our proposed Med-VLFM ranked 1^{st} on most of the automatic evaluation metrics, including BERTScore, ROUGE-1, BLEU-1, BLEURT, METEOR, CIDEr and RefCLIPScore, in the ImageCLEFmedical 2024 caption prediction challenge. Our code and models are available at https://openi.pcl.ac.cn/OpenMedIA/PCLmed24.

Keywords

Medical Image Captioning, Vision-Language Models, Medical Vision-Langauge Foundation Models, Parameter-Efficient Fine-Tuning, Vision Encoder Ensembling, Modality-Aware Adaptation

1. Introduction

Medical report generation (MRG), the main application of image captioning [1] in the medical domain, stands as a pivotal task in healthcare, aiming to automatically produce precise and coherent reports delineating the impressions and observations derived from medical images [2, 3, 4]. High-quality cross-modal annotations, comprising meticulously paired medical image-report datasets, are essential for the success of MRG. Despite the emergence of a few open-source datasets fostering research in this domain [5, 6, 7], their limited scale poses significant challenges for the development of deep and expansive neural architectures.

Recent advancements in vision and language applications have led to the rise of large-scale pre-trained models, termed foundation models (FMs), designed for general applicability, and demonstrate versatility across various tasks, owing to prompt engineering or parameter-efficient fine-tuning (PEFT). The scalability of both model and data enables FMs to acquire emergent capabilities, empowering them to tackle tasks previously deemed challenging for smaller models [8, 9, 10]. In the pursuit of effective MRG, researchers have explored methodologies to harness the capabilities of FMs while addressing the scarcity of labelled medical image-report pairs. One notable approach is the BLIP-2 architecture [11], a state-of-the-art vision-language pre-training methodology that facilitates knowledge transfer from single-modality vision and language foundation models.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*]Corresponding author.

yangbang@pku.edu.cn (B. Yang); yuy@pcl.ac.cn (Y. Yu); zouyx@pku.edu.cn (Y. Zou); zhangt02@pcl.ac.cn (T. Zhang)
https://web.pkusz.edu.cn/adsp/ (Y. Zou)
0000-0003-2019-0377 (B. Yang); 0000-0002-8838-4963 (T. Zhang)

In this study, we present a novel approach to address the challenges in medical image captioning and report generation using medical vision-language foundation models (Med-VLFMs). Drawing inspiration from the BLIP-2 and its adaptations in the medical domain [11, 12, 13], our Med-VLFM incorporates a lightweight query Transformer (Q-Former) to connect three foundation models (FMs): an ensemble of EVA ViT-g [8] and BiomedCLIP [14] serving as the vision encoder, and Pangu-α [15] as the language decoder. Our proposed Med-VLFM introduces two innovative techniques: vision encoder ensembling (VEE) and modality-aware adaptation (MAA). VEE combines general and medical specialist vision foundation models to create a more comprehensive representation of medical images. MAA calibrates vision and text features using small sets of trainable parameters in LLMs. Through experimentation on the ImageCLEFmedical 2024 caption prediction challenge [16, 17], our Med-VLFM ranked 1^{st} on most of the automatic evaluation metrics, including BERTScore, ROUGE-1, BLEU-1, BLEURT, METEOR, CIDEr and RefCLIPScore, demonstrating the effectiveness of VEE and MAA in enhancing medical image captioning and report generation.

2. Related Works

Medical/Radiology Report Generation Motivated by the rapid development of image captioning [1, 18, 19, 20, 21], the field of medical report generation has seen significant research interest in recent years. Different from pure-text scenarios like chatting with patients [22, 23, 24] and discharge instruction generation [25], medical report generation needs to "translate" medical images into detailed reports. As such, one line of research focus on improving the cross-modal alignment between medical images and reports, which is usually achieved by reinforcement learning [3, 26], contrastive learning [27, 28], well-designed modules like hierarchical attention [29] and memory [30]. Given that generating accurate reports requires domain expertise, another line of research opts to provide models with effective priors through retrieval [3, 31, 32] or augment models with knowledge [28, 31, 33]. However, the language models used in these medical report generation work are typically shallow and may lack the capacity to capture the nuances of context and execute complex reasoning.

Medical Vision-Language Foundation Models Recent advancements in conversational AI have shown promise in aiding biomedical practitioners. LLaVA-Med [34] proposes an efficient approach to training a vision-language conversational assistant for answering biomedical image research questions. Med-PaLM [35] provides high-quality answers to medical inquiries, while R2GenGPT [36] enhances Radiology Report Generation by aligning visual features with language model embeddings. Additionally, XrayGPT [37] introduces a conversational medical vision-language model for analyzing chest radiographs. While these models signify significant progress in multimodal conversational AI for the medical domain, their efficacy relies heavily on the quality and quantity of paired training samples. As such, Med-MLLM [38] learns radiology representations from unlabelled data to quickly deploy tools for rapid response to rare diseases. Despite the above efforts, the potential benefits of incorporating multiple vision models have yet to be explored.

Parameter-Efficient Fine-Tuning With the proliferation of foundation models (FMs) [9, 10], efficiently adapting FMs to a specific task becomes a research hotspot. One effective technique is *prompt engineering* [39], which aims to affect the behaviors of language FMs by providing them with a textual template filled with task-related priors [40, 41], demonstrations of several examples [42, 43], or a chain of thoughts [44, 45, 46]. Alongside prompt engineering, parameter-efficient fine-tuning (PEFT) has also emerged as a popular technique to influence the intermediate hidden states and final responses of FMs. In implementation, PEFT either introduces lightweight components, e.g., Adapter [47], continuous prompts [41, 48], and LoRA [49], vectors that scale the inner activations [50], into FMs, or adapts a small portion of inherent weights of FMs [51]. Recent practices utilizing these two techniques have demonstrated the effectiveness of adapting general-purpose FMs to the medical domain [52, 53, 54]. Distinct from these practices, our work focuses on medical report generation.

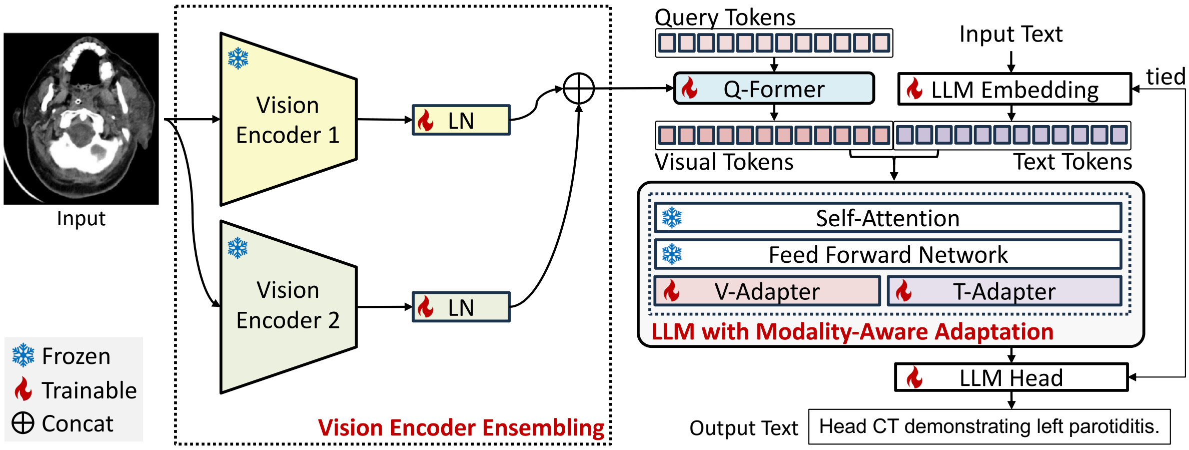


Figure 1: Overview of our proposed Med-VLFM for medical/radiology report generation.

3. Approach

Overview As shown in Figure 1, our proposed Med-VLFM for medical/radiology report generation comprises four major components: vision encoders parameterized by $\Theta_v = \{\theta_v^1, \theta_v^2\}$, a query Transformer (Q-Former) parameterized by θ_q, a LLM parameterized by θ_t, and modality-aware adaptation parameterized by $\Delta = \{\delta_v, \delta_t\}$. Given the medical image I and the target caption/report $\mathbf{y} = \{y_1, y_2, \ldots, y_N\}$, our model minimizes the following negative log-likelihood:

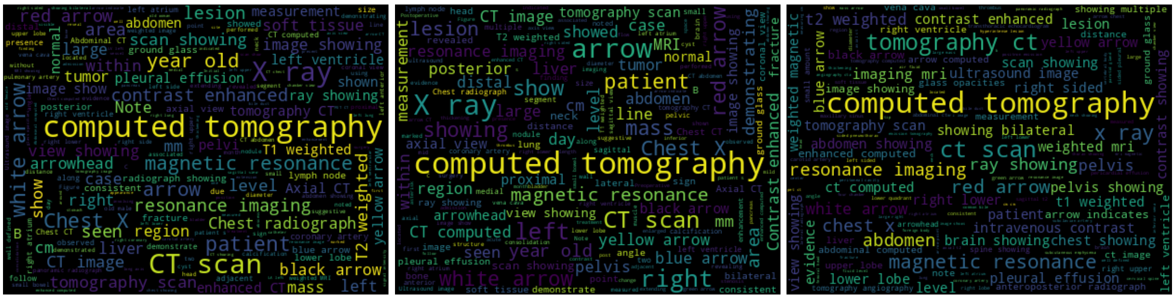
$$\mathcal{L} = -\sum_{n=1}^{N} \log p(y_n|\mathbf{y_{<n}}, I; \{\Theta_v, \theta_q, \theta_t, \Delta\}). \tag{1}$$

Next, we elaborate on each component of our model.

Vision Encoder Ensembling We consider general-purpose and specialist vision encoders to produce comprehensive visual representations and thus adopt EVA-ViT-g [8] pre-trained on 29.6 million natural images and BioMedCLIP [14] pre-trained on 15 million medical images crawled from the scientific publications in PubMed Central[1]. We use the fine-tuned weights from our previous work [12] for EVA-ViT-g and the official weights for BioMedCLIP. Following BLIP-2 [11], we remove the last layer of these encoders and use the second last layer's output features. To stabilize training, a layer normalization layer [55] is added to the end of each vision encoder. As different vision encoders may have their distinct image processing pipelines, we adhere to their original settings and feed images of resolution 364 × 364 and 224 × 224 into EVA-ViT-g and BioMedCLIP, resulting in features $\boldsymbol{R}_1 \in \mathbb{R}^{K_1 \times d_1}$ and $\boldsymbol{R}_2 \in \mathbb{R}^{K_2 \times d_2}$, respectively. For fusion, we apply the non-parametric bicubic interpolation operation on $\boldsymbol{R}_2$ to obtain $\boldsymbol{R}'_2 \in \mathbb{R}^{K_1 \times d_2}$ first and then concatenate $\boldsymbol{R}_1$ and $\boldsymbol{R}'_2$ along the channel dimension to attain the final visual features $\boldsymbol{R} \in \mathbb{R}^{K_1 \times (d_1+d_2)}$. In our case, $K_1 = 676, K_2 = 196, d_1 = 1408, d_2 = 768$.

Query Transformer (Q-Former) Directly feeding $\boldsymbol{R}$ into the subsequent language modeling process will introduce redundancy and lead to high memory and computational costs due to the quadratic nature of the standard self-attention mechanism [56]. Thus, we adopt a Q-former with L ($L \ll K_1$) learnable query tokens to aggregate $\boldsymbol{R}$ into visual tokens $\boldsymbol{V} \in \mathbb{R}^{L \times d_{\text{llm}}}$. In practice, we follow the setting of [13]: Q-Former is a randomly initialized BERT-like encoder [57] and has $L = 32$ query tokens, 6 Transformer blocks with hidden size $d_q = 768$ and cross-attention layers inserted at a frequency of 2, and a linear projection layer that maps d_q to d_{llm}.

[1] https://www.ncbi.nlm.nih.gov/pmc/



(a) (b) (c)

Figure 2: Word clouds of ground-truth reports from the training set (a) and the validation set (b). In (c), we show the word cloud of captions generated by our best-performed model on the testing set.

Large Language Model (LLM) We use Pangu-α [58] for language modeling according to our computation budget. In particular, we replace its vocabulary with that of OPT-2.7B [59], since Pangu-α was trained mainly with Chinese words. To ensure the alignment between randomly initialized token embeddings and the other weights of Pangu-α, we train token embeddings while keeping the rest of the LLM parameters frozen. As Pangu-α is a decoder-only Transformer, we prefix its original input sequence, i.e., text tokens $\boldsymbol{T} \in \mathbb{R}^{N \times d_{\text{llm}}}$, with $\boldsymbol{V}$ following common practices [36, 60] to achieve vision-grounded medical report generation. The resulting LLM's input sequence is denoted as $\boldsymbol{H} = [\boldsymbol{V}; \boldsymbol{T}] \in \mathbb{R}^{(L+N) \times d_{\text{llm}}}$. We note that using more advanced or medical-related LLMs [61, 62, 63] may yield better performance. We leave this exploration to our future study.

Modality-Aware Adaptation Let's assume the LLM has J blocks, each of which is parameterized by $\theta_t^{(j)}$ ($j \in [1, J]$), and denote $\boldsymbol{H}^{(j)} = [\boldsymbol{V}^{(j)}; \boldsymbol{T}^{(j)}]$ as the original output of the j-th block, we can compute $\boldsymbol{H}^{(j)}$ as follows:

$$\boldsymbol{H}^{(j)} = \text{LLMBlock}(\boldsymbol{H}^{(j-1)}; \theta_t^{(j)}), \tag{2}$$

where $\boldsymbol{H}^{(0)} = \boldsymbol{H}$. There are two incoming problems: (1) $\theta_t^{(j)}$ is optimized for the text-only modality, which can be unsuitable for the mixed-modal input due to the modality gap [64] and (2) using the same set of parameters to learn multimodal representations may limit the collaboration of vision and text modalities [65]. Therefore, we propose modality-aware adaptation (MAA) to mitigate the above two problems. In particular, MAA introduces light-weight adaptation modules independent of blocks and modalities, i.e., the newly added parameters can be defined as $\Delta^{(j)} = \{\delta_v^{(j)}, \delta_t^{(j)}\}$. In this work, we consider a simple implementation of MAA: putting adaptation modules right after each block to calibrate $\boldsymbol{H}^{(j)}$. So Eq. (2) is modified as follows:

$$\begin{aligned} \boldsymbol{H}^{(j)} &= \text{LLMBlock}(\overline{\boldsymbol{H}}^{(j-1)}; \theta_t^{(j)}), \\ \overline{\boldsymbol{H}}^{(j)} &= \text{Adaptation}(\boldsymbol{H}^{(j)}; \Delta^{(j)}) \\ &= [\boldsymbol{V}^{(i)} + \text{Adapter}(\boldsymbol{V}^{(i)}; \delta_v^{(j)}); \boldsymbol{T}^{(i)} + \text{Adapter}(\boldsymbol{T}^{(i)}; \delta_t^{(j)})], \end{aligned} \tag{3}$$

where $\overline{\boldsymbol{H}}^{(0)} = \boldsymbol{H}$, pre-trained weights $\theta_t^{(j)}$ are kept frozen, $[;]$ denotes concatenation along the sequence dimension, and $\text{Adapter}(\cdot)$ is a bottle-neck MLP as in [47], i.e., $\text{Adapter}(x) = \boldsymbol{W}_{\text{up}}(\sigma \boldsymbol{W}_{\text{down}}(x))$. In the implementation, we set the non-linearity activation function σ as GELU [66] and the bottleneck size of adapters to 64. Moreover, inspired by the zero-initialization technique [49], we randomly initialize $\boldsymbol{W}_{\text{up}}$ but initialize $\boldsymbol{W}_{\text{down}}$ with all zeros, so that the intermediate and output features of LLM can be adapted smoothly.

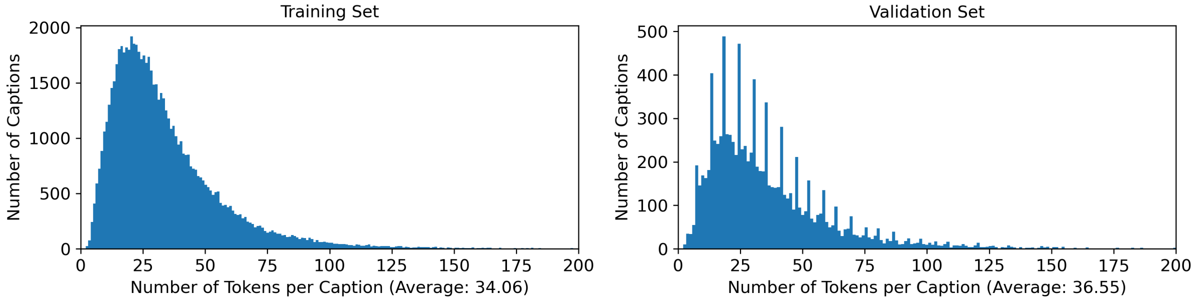


Figure 3: Histograms of the number of tokens per caption. We use GPT-2's tokenizer [67] to split captions into tokens. Roughly 99% captions contain less than 128 tokens.

4. Experiments

4.1. Experimental Setups

Dataset The development dataset of the ImageCLEFmedical 2024 caption prediction challenge [17] is ROCOv2 [7], which is an updated and extended version of the Radiology Objects in COntext (ROCO) dataset [68]. ROCOv2 provides 70,108, 9,972, and 17,237 radiology images for training, validation, and testing respectively. Images originating from biomedical articles are annotated with one medical caption each. In Figure 2 (a) and (b), we visualize the word clouds of ground-truth reports from the training and validation sets. As we can observe, there are many computed tomography (CT) and X-ray images in the dataset. Besides, some common expressions like "black arrow" indicate that humans have marked a large portion of images.

Metrics As noted in the guidelines of the competition[2], the major and secondary metrics used for the challenge are BERTScore [69] and ROUGE-1 [70]. Specifically, BERTScore is a model-based metric that calculates the semantic similarity of two sentences. ROUGE-1 measures the number of matching unigrams between a model-generated text and a reference. Besides, we also report BLEU [71], METEOR [72], CIDEr [73], BLEURT [74], CLIPScore and RefCLIPScore [75] to comprehensively evaluate the effectiveness of our proposals. For all metrics, higher is better.

Comparing Model We treat our last year's solution [12] (abbreviated as PCLmed-23) as a baseline. It adopts EVA-ViT-g [8] as the vision encoder, v1.0 ChatGLM-6B[3] [76] as the decoder, and adapts ChatGLM-6B with P-Tuning [48] (please see the original paper for more details). We use PCLmed-23's original hyper-parameters and directly evaluate PCLmed-23 on the validation set of ROCOv2. Note that although we do not train PCLmed-23 on ROCOv2 (this year's data), there is a significant overlap between last year's and this year's training data.

Image and Text Processing During training, we process images with random resized cropping with ratios falling into $[0.9, 1.0]$. The cropped images are resized to the maximum resolution required by vision encoders, i.e., 364×364 in this work. We apply bilinear downsampling on the same cropped images for vision encoders with lower-resolution images as inputs. Unlike PCLmed-23, we keep all symbols in texts this year and truncate them into a maximum length of 128 based on the histograms shown in Fig. 3.

Model Settings Most model details have been introduced in Section 3. Unlike PCLmed-23, we instruct the LLM with an empty text prompt since we found it is slightly better than the text prompt

[2] https://www.imageclef.org/2024/medical/caption
[3] https://github.com/THUDM/ChatGLM-6B

Table 1
BERTScore (main metric) in the ImageCLEFmedical 2024 caption prediction challenge. †: EVA-ViT-G has been fine-tuned on the training data of the last year's challenge. *: We do not train PCLmed-23 on data from this year, but there is a significant overlap between last year's and this year's training data.

Model	**Vision Encoder**	**LLM Type**	**LLM Adaptation**	**#Parameters**		**BERTScore**	
				Total	Trainable	Validation	Test
PCLmed-23	EVA-ViT-G†	ChatGLM-6B	Modality-Agnostic (P-Tuning)	7.3B	N/A*	0.610308	-
#1	BioMedCLIP	Pangu-α	None	2.8B	180.3M	0.629966	-
#2	EVA-ViT-G†	Pangu-α	None	3.8B	183.3M	0.632418	0.622711
#3	#1 + #2	Pangu-α	None	3.8B	186.8M	0.633252	0.623535
#4	#1 + #2	Pangu-α	Modality-Agnostic	3.8B	192.1M	0.637506	-
#5 (Ours)	#1 + #2	Pangu-α	Modality-Aware	3.8B	197.4M	**0.638812**	**0.629913**

Table 2
Full validation performance in the ImageCLEFmedical 2024 caption prediction challenge.

Model	**Main Metrics**		**Other Metrics**								
	BERTScore	ROUGE-1	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	CIDEr	BLEURT	CLIPScore	RefCLIPScore
PCLmed-23	0.610308	0.257961	0.231894	0.127473	0.068828	0.039036	0.101455	0.196403	0.330486	0.813100	0.809721
#1	0.629966	0.299996	0.283482	0.167054	0.100009	0.062129	0.117641	0.249127	0.347528	**0.826615**	0.817631
#2	0.632418	0.303293	0.284882	0.168868	0.102936	0.064572	0.118730	0.267878	0.346301	0.824120	0.819503
#3	0.633252	**0.305291**	0.287658	0.171318	0.104373	0.065453	0.120168	0.269210	0.348228	0.825405	0.820024
#4	0.637506	0.302669	0.286492	0.171013	0.104543	0.066096	0.120499	0.276311	0.347248	0.825275	0.819929
#5 (Ours)	**0.638812**	0.304166	**0.289269**	**0.172753**	**0.105881**	**0.066715**	**0.121668**	**0.278013**	**0.348230**	0.825773	**0.820519**

like "generate a medical report for the input image" in our preliminary experiments. Besides, we insert the adaptation modules (i.e., Adapter) into the LLM at a frequency of 2.

Hyper-Parameters During training, we use AdamW [77] and L2 weight decay of 0.05 to train models with 32 samples per batch for 10 epochs. The learning rate (lr) is increased to 1e-4 in 1,000 warm-up steps and follows a cosine annealing scheduler. We train token embeddings with $0.1 \times lr$ to stabilize training. During evaluation, we use the beam search decoding algorithm with a beam size of 3 to generate medical captions and force the model to generate at least 8 tokens and up to 64 tokens. We set the repetition penalty to 2.5 to avoid duplication and the length penalty to 2.0 to encourage generating longer captions.

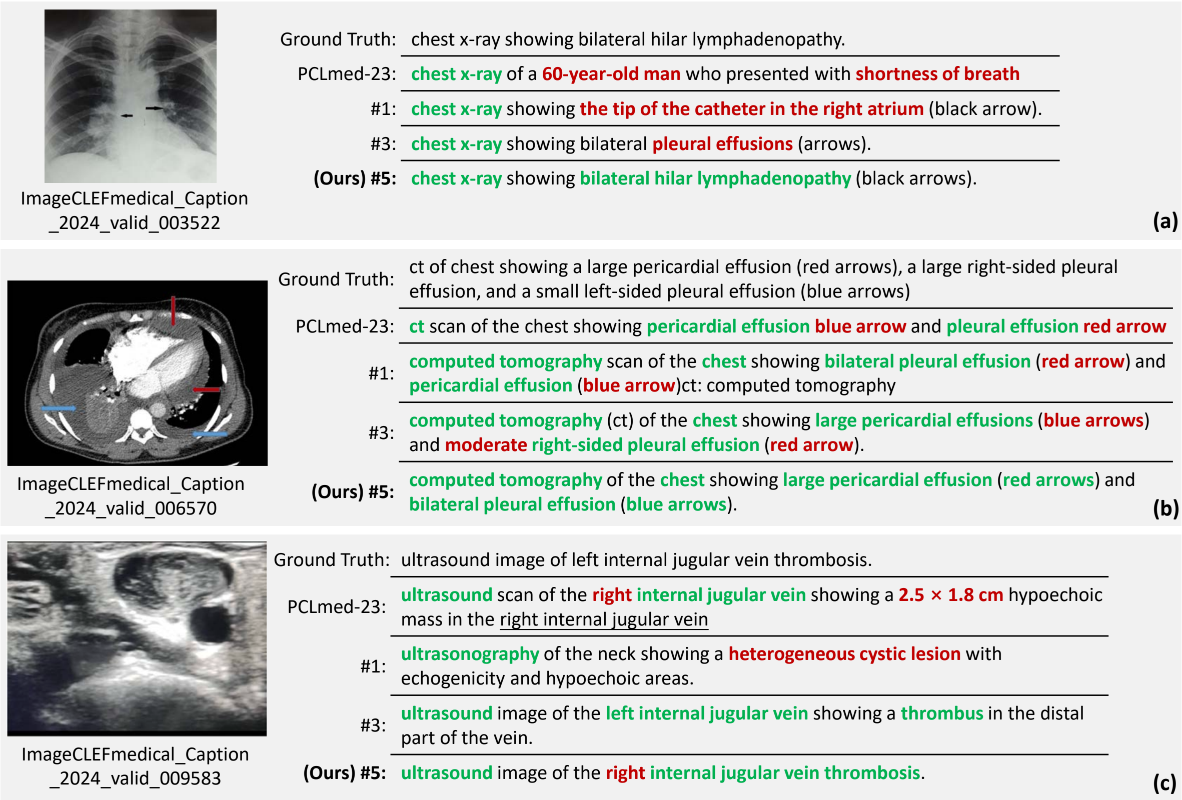
4.2. Quantitative Results

We submit three runs to the ImageCLEFmedical 2024 caption prediction challenge, which correspond to #{2, 3, 5} in Tables 1 and 2. In addition, we also run some ablations locally to highlight the effectiveness of our proposals. Based on quantitative results in Tables 1 and 2, We have the following observations.

- Comparing #{1, 2, 3}, we can see that ensembling BioMedCLIP and EVA-ViT-G boosts performance on all metrics except CLIPScore. This is because CLIPScore is a reference-free metric, meaning that the metric score could be biased by the pre-trained knowledge of CLIP [78]. Instead, RefCLIPScore considers the semantic similarity between references and predictions and #3 outperforms #{1, 2}. From another perspective, the overall performance improvements suggest that different pre-trained vision models may have learned complementary visual features and our adopted ensembling/fusion strategy, i.e., concatenating visual features along the channel dimension, helps produce the holistic representations of medical images.
- Comparing #{3, 4}, we can observe that fine-tuning LLM with adapters regardless of vision and text modalities may suffer from performance degradation. For example, #4 performs worse than

- #3 on 6 out of 11 metrics listed in Table 2.
- Comparing #{4, 5}, we can see consistent improvements for all metrics if we adapt vision and text features in LLM independently. This suggests that we should allocate modality-specific parameters within LLM to alleviate modality interference and boost modality collaboration.

In short, the above observations verify the effectiveness of our proposed vision encoder ensembling and modality-aware adaptation. Our final model (#5) surpasses PCLmed-23 and generally boosts performance by a large margin compared with #{1, 2}. Nonetheless, there still exist many potential improvement directions of our proposals, e.g., (1) designing a more advanced ensembling/fusion strategy to unify the intelligence of different vision foundation models for medical tasks and (2) exploring a more reasonable way to insert/allocate modality-specific parameters within LLM.



ImageCLEFmedical_Caption_2024_valid_003522

Ground Truth:	chest x-ray showing bilateral hilar lymphadenopathy.
PCLmed-23:	chest x-ray of a 60-year-old man who presented with shortness of breath
#1:	chest x-ray showing the tip of the catheter in the right atrium (black arrow).
#3:	chest x-ray showing bilateral pleural effusions (arrows).
(Ours) #5:	chest x-ray showing bilateral hilar lymphadenopathy (black arrows).

ImageCLEFmedical_Caption_2024_valid_006570

Ground Truth:	ct of chest showing a large pericardial effusion (red arrows), a large right-sided pleural effusion, and a small left-sided pleural effusion (blue arrows)
PCLmed-23:	ct scan of the chest showing pericardial effusion blue arrow and pleural effusion red arrow
#1:	computed tomography scan of the chest showing bilateral pleural effusion (red arrow) and pericardial effusion (blue arrow)ct: computed tomography
#3:	computed tomography (ct) of the chest showing large pericardial effusions (blue arrows) and moderate right-sided pleural effusion (red arrow).
(Ours) #5:	computed tomography of the chest showing large pericardial effusion (red arrows) and bilateral pleural effusion (blue arrows).

ImageCLEFmedical_Caption_2024_valid_009583

Ground Truth:	ultrasound image of left internal jugular vein thrombosis.
PCLmed-23:	ultrasound scan of the right internal jugular vein showing a 2.5 × 1.8 cm hypoechoic mass in the right internal jugular vein
#1:	ultrasonography of the neck showing a heterogeneous cystic lesion with echogenicity and hypoechoic areas.
#3:	ultrasound image of the left internal jugular vein showing a thrombus in the distal part of the vein.
(Ours) #5:	ultrasound image of the right internal jugular vein thrombosis.

Figure 4: Qualitative examples on the validation set. We mark accurate keywords in green, wrong details in red, and underline repeated content. Image sources from top to bottom: CC BY-NC [Chauhan et al. (2021)], CC BY [Muacevic et al. (2022)], and CC BY [Laaribi et al. (2021)].

4.3. Qualitative Analysis

In Figure 4, we visualize three qualitative examples, where ground-truth captions and captions generated by different models are presented. We have the following observations.

- All models can precisely identify different imaging modalities, i.e., X-Ray in (a), Computed Tomography in (b), and Ultrasonography in (c).
- In Figure 4 (b), we can see that all models can predict the existence of “pericardial effusion”, “pleural effusion”, and blue and red arrows. However, PCLmed-23 and #{1, 3} fail to ground the abnormalities to the images, i.e., they mistakenly relate “pericardial effusion” to blue arrows and “pleural effusion” to red arrows. This suggests the importance of improving Med-VLFMs’ grounding abilities.

- Benefited from the proposed vision encoder ensembling and modality-aware adaptation, our model (#5) captures the abnormality in medical images more accurately, e.g., “bilateral hilar lymphadenopathy” in (a) and “internal jugular vein thrombosis” in (c).
- Although our model (#5) generally performs the best, it still makes simple mistakes, i.e., “right” in (c). There are several possible solutions for this problem, e.g., (1) enlarging the image resolution [79] and (2) incorporating positional embeddings into the Q-Former-like connection module [62].

5. Conclusion

In this study, we propose a parameter-efficient training pipeline for medical image captioning and report generation with single-modality pre-trained vision FMs and LLMs. By introducing vision encoder ensembling and modality-aware adaptation, our method leverages the merits of both general and medical vision models and calibrates vision and text features in the LLM. Our Med-VLFM achieved 1st place in the ImageCLEFmedical 2024 caption prediction challenge, excelling across multiple evaluation metrics, including BERTScore, ROUGE-1, BLEU-1, BLEURT, METEOR, CIDEr, and RefCLIPScore. This success highlights the potential of our approach to significantly enhance medical AI solutions and support radiologists in their work. Source codes with model weights are available at OpenMedIA[4] [80].

Acknowledgments

This work is partially supported by the Guangdong Provincial Science and Technology Department under Grant No. 2024A1515010245, in part by the National Natural Science Foundation of the National Natural Science Foundation of China (grant No. U21A20523 and 62176008), and in part by the Major Key Project PCL under grant No. PCL2023A09. The computing resources of Pengcheng Cloudbrain are used in this research. We acknowledge the support provided by OpenI Community[5].

References

[1] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: ICML, PMLR, 2015, pp. 2048–2057.

[2] B. Jing, P. Xie, E. Xing, On the automatic generation of medical imaging reports, in: ACL, 2018, pp. 2577–2586.

[3] Y. Li, X. Liang, Z. Hu, E. P. Xing, Hybrid retrieval-generation reinforced agent for medical image report generation, in: NeurIPS, volume 31, 2018.

[4] Z. Chen, Y. Song, T.-H. Chang, X. Wan, Generating radiology reports via memory-driven Transformer, in: EMNLP, 2020, pp. 1439–1449.

[5] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, C. J. McDonald, Preparing a collection of radiology examinations for distribution and retrieval, Journal of the American Medical Informatics Association 23 (2016) 304–310. doi:10.1093/jamia/ocv080.

[6] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Scientific Data 6 (2019) 317. doi:10.1038/s41597-019-0322-0.

[7] J. Rückert, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, S. Koitka, O. Pelka, A. B. Abacha, A. G. S. de Herrera, H. Müller, P. A. Horn, F. Nensa, C. M. Friedrich, ROCOv2: Radiology objects in context version 2, an updated multimodal image dataset, Scientific Data (2024). doi:10.1038/s41597-024-03496-6.

[4] https://openi.pcl.ac.cn/OpenMedIA
[5] https://git.openi.org.cn

[8] Y. Fang, W. Wang, B. Xie, Q. Sun, L. Wu, X. Wang, T. Huang, X. Wang, Y. Cao, EVA: Exploring the limits of masked visual representation learning at scale, in: CVPR, 2023, pp. 19358–19369.

[9] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, J. Gao, Large language models: A survey, arXiv preprint arXiv:2402.06196 (2024).

[10] W. Yang, M. Liu, Z. Wang, S. Liu, Foundation models meet visualizations: Challenges and opportunities, Computational Visual Media 10 (2024) 399–424. doi:10.1007/s41095-023-0393-x.

[11] J. Li, D. Li, S. Savarese, S. Hoi, BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: ICML, 2023, pp. 19730–19742.

[12] B. Yang, A. Raza, Y. Zou, T. Zhang, PCLmed at ImageCLEFmedical 2023: Customizing general-purpose foundation models for medical report generation, in: CLEF2023 Working Notes, 2023.

[13] S. Wu, B. Yang, Z. Ye, H. Wang, H. Zheng, T. Zhang, MAKEN: Improving medical report generation with adapter tuning and knowledge enhancement in vision-language foundation models, in: ISBI, 2024.

[14] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, et al., BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, arXiv preprint arXiv:2303.00915 (2023).

[15] W. Zeng, X. Ren, T. Su, H. Wang, Y. Liao, Z. Wang, X. Jiang, Z. Yang, K. Wang, X. Zhang, et al., Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation, arXiv preprint arXiv:2104.12369 (2021).

[16] B. Ionescu, H. Müller, A. Drăgulinescu, J. Rückert, A. Ben Abacha, A. Garcıa Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, C. S. Schmidt, T. M. G. Pakull, H. Damm, B. Bracke, C. M. Friedrich, A. Andrei, Y. Prokopchuk, D. Karpenka, A. Radzhabov, V. Kovalev, C. Macaire, D. Schwab, B. Lecouteux, E. Esperança-Rodier, W. Yim, Y. Fu, Z. Sun, M. Yetisgen, F. Xia, S. A. Hicks, M. A. Riegler, V. Thambawita, A. Storås, P. Halvorsen, M. Heinrich, J. Kiesel, M. Potthast, B. Stein, Overview of ImageCLEF 2024: Multimedia retrieval in medical applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024), Springer Lecture Notes in Computer Science LNCS, Grenoble, France, 2024.

[17] J. Rückert, A. Ben Abacha, A. G. Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, B. Bracke, H. Damm, T. M. G. Pakull, C. S. Schmidt, H. Müller, C. M. Friedrich, Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[18] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, H. Laga, A comprehensive survey of deep learning for image captioning, ACM Computing Surveys 51 (2019) 1–36. doi:10.1145/3295748.

[19] F. Liu, Y. Liu, X. Ren, X. He, X. Sun, Aligning visual regions and textual concepts for semantic-grounded image representations, in: NeurIPS, volume 32, 2019.

[20] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, L. Wang, Scaling up vision-language pre-training for image captioning, in: CVPR, 2022, pp. 17980–17989.

[21] B. Yang, F. Liu, X. Wu, Y. Wang, X. Sun, Y. Zou, MultiCapCLIP: Auto-encoding prompts for zero-shot multilingual visual captioning, in: ACL, 2023, pp. 11908–11922.

[22] Y. Li, Z. Li, K. Zhang, R. Dan, S. Jiang, Y. Zhang, ChatDoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, Cureus (2023). doi:10.7759/cureus.40895.

[23] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, D. S. W. Ting, Large language models in medicine, Nature Medicine 29 (2023) 1930–1940. doi:10.1038/s41591-023-02448-8.

[24] S. Liu, A. B. McCoy, A. P. Wright, B. Carew, J. Z. Genkins, S. S. Huang, J. F. Peterson, B. Steitz, A. Wright, Leveraging large language models for generating responses to patient messages—a subjective analysis, Journal of the American Medical Informatics Association 31 (2024) 1367–1379. doi:10.1093/jamia/ocae052.

[25] F. Liu, B. Yang, C. You, X. Wu, S. Ge, Z. Liu, X. Sun, Y. Yang, D. Clifton, Retrieve, reason, and refine: Generating accurate and faithful patient instructions, in: Advances in Neural Information

Processing Systems, volume 35, 2022, pp. 18864–18877.

[26] H. Qin, Y. Song, Reinforced cross-modal alignment for radiology report generation, in: ACL Findings, 2022, pp. 448–458.

[27] Y. Li, B. Yang, X. Cheng, Z. Zhu, H. Li, Y. Zou, Unify, align and refine: Multi-level semantic alignment for radiology report generation, in: ICCV, 2023, pp. 2863–2874.

[28] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, X. Chang, Dynamic graph enhanced contrastive learning for chest x-ray report generation, in: CVPR, 2023, pp. 3334–3343.

[29] D. You, F. Liu, S. Ge, X. Xie, J. Zhang, X. Wu, Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation, in: MICCAI, 2021, pp. 72–82.

[30] Z. Chen, Y. Shen, Y. Song, X. Wan, Cross-modal memory networks for radiology report generation, in: ACL, 2021, pp. 5904–5914.

[31] C. Y. Li, X. Liang, Z. Hu, E. P. Xing, Knowledge-driven encode, retrieve, paraphrase for medical image report generation, AAAI 33 (2019) 6666–6673.

[32] F. Liu, X. Wu, S. Ge, W. Fan, Y. Zou, Exploring and distilling posterior and prior knowledge for radiology report generation, in: CVPR, 2021, pp. 13753–13762.

[33] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, D. Xu, When radiology report generation meets knowledge graph, in: AAAI, volume 34, 2020, pp. 12910–12917.

[34] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, J. Gao, LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day, in: NeurIPS, volume 36, 2023, pp. 28541–28564.

[35] T. Tu, S. Azizi, D. Driess, M. Schaekermann, M. Amin, P.-C. Chang, A. Carroll, C. Lau, R. Tanno, I. Ktena, A. Palepu, B. Mustafa, A. Chowdhery, Y. Liu, S. Kornblith, D. Fleet, P. Mansfield, S. Prakash, R. Wong, S. Virmani, C. Semturs, S. S. Mahdavi, B. Green, E. Dominowska, B. A. y Arcas, J. Barral, D. Webster, G. S. Corrado, Y. Matias, K. Singhal, P. Florence, A. Karthikesalingam, V. Natarajan, Towards generalist biomedical ai, NEJM AI 1 (2024) AIoa2300138. doi:10.1056/AIoa2300138.

[36] Z. Wang, L. Liu, L. Wang, L. Zhou, R2GenGPT: Radiology report generation with frozen llms, Meta-Radiology 1 (2023) 100033. doi:10.1016/j.metrad.2023.100033.

[37] O. Thawkar, A. Shaker, S. S. Mullappilly, H. Cholakkal, R. M. Anwer, S. Khan, J. Laaksonen, F. S. Khan, XrayGPT: Chest radiographs summarization using medical vision-language models, 2023. arXiv:2306.07971.

[38] F. Liu, T. Zhu, X. Wu, B. Yang, C. You, C. Wang, L. Lu, Z. Liu, Y. Zheng, X. Sun, Y. Yang, L. Clifton, D. A. Clifton, A medical multimodal large language model for future pandemics, npj Digital Medicine 6 (2023) 1–15. doi:10.1038/s41746-023-00952-2.

[39] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35. doi:10.1145/3560815.

[40] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text Transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[41] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, in: ACL, 2021, pp. 4582–4597.

[42] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: NeurIPS, 2020, pp. 1877–1901.

[43] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al., Flamingo: a visual language model for few-shot learning, in: NeurIPS, 2022, pp. 23716–23736.

[44] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, et al., Chain-of-thought prompting elicits reasoning in large language models, in: NeurIPS, 2022.

[45] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, in: NeurIPS, 2022, pp. 22199–22213.

[46] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, et al., PaLM: Scaling language modeling with pathways, in: ICLR, 2023,

pp. 1–33.

[47] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: ICML, PMLR, 2019, pp. 2790–2799.

[48] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, J. Tang, P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks, in: ACL, 2022, pp. 61–68.

[49] E. J. Hu, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al., LoRA: Low-rank adaptation of large language models, in: ICLR, 2022, pp. 1–13.

[50] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, C. A. Raffel, Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, in: NeurIPS, volume 35, 2022, pp. 1950–1965.

[51] E. Ben Zaken, Y. Goldberg, S. Ravfogel, Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, in: ACL, 2022, pp. 1–9.

[52] V. Liévin, C. E. Hother, O. Winther, Can large language models reason about medical questions?, Patterns 5 (2024). doi:10.1016/j.patter.2024.100943.

[53] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., Large language models encode clinical knowledge, Nature 620 (2022) 172 – 180. doi:10.1038/s41586-023-06291-2.

[54] H. Nori, N. King, S. M. McKinney, D. Carignan, E. Horvitz, Capabilities of GPT-4 on medical challenge problems (2023). arXiv:2303.13375.

[55] J. L. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, 2016. arXiv:1607.06450.

[56] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017).

[57] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional Transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), NAACL-HLT, 2019, pp. 4171–4186.

[58] W. Zeng, X. Ren, T. Su, H. Wang, Y. Liao, Z. Wang, X. Jiang, Z. Yang, K. Wang, X. Zhang, C. Li, Z. Gong, Y. Yao, X. Huang, J. Wang, J. Yu, Q. Guo, Y. Yu, Y. Zhang, J. Wang, H. Tao, D. Yan, Z. Yi, F. Peng, F. Jiang, H. Zhang, L. Deng, Y. Zhang, Z. Lin, C. Zhang, S. Zhang, M. Guo, S. Gu, G. Fan, Y. Wang, X. Jin, Q. Liu, Y. Tian, Pangu-α: Large-scale autoregressive pretrained chinese language models with auto-parallel computation, 2021. arXiv:2104.12369v1.

[59] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, et al., OPT: Open pre-trained Transformer language models, 2022. arXiv:2205.01068v4.

[60] C. Liu, Y. Tian, W. Chen, Y. Song, Y. Zhang, Bootstrapping large language models for radiology report generation, in: AAAI, volume 38, 2024, pp. 18635–18643.

[61] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al., LLaMA: Open and efficient foundation language models, 2023. arXiv:2302.13971.

[62] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. arXiv:2308.12966v3.

[63] C. Wu, W. Lin, X. Zhang, Y. Zhang, Y. Wang, W. Xie, Pmc-llama: Towards building open-source language models for medicine, Journal of the American Medical Informatics Association (2024). doi:10.1093/jamia/ocae045.

[64] V. W. Liang, Y. Zhang, Y. Kwon, S. Yeung, J. Y. Zou, Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning, in: NeurIPS, volume 35, 2022, pp. 17612–17625.

[65] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, F. Huang, J. Zhou, mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration, in: CVPR, 2024, pp. 13040–13051.

[66] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), 2023. arXiv:1606.08415v5.

[67] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI blog (2019).

[68] O. Pelka, S. Koitka, J. Rückert, F. Nensa, C. 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, Springer, 2018, pp. 180–189.

[69] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, in: ICLR, 2019.

[70] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[71] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, BLEU: A method for automatic evaluation of machine translation, in: ACL, 2002, pp. 311–318.

[72] S. Banerjee, A. Lavie, METEOR: An automatic metric for mt evaluation with improved correlation with human judgments, in: ACL Workshop, 2005, pp. 65–72.

[73] R. Vedantam, C. Lawrence Zitnick, D. Parikh, CIDEr: Consensus-based image description evaluation, in: CVPR, 2015, pp. 4566–4575.

[74] T. Sellam, D. Das, A. Parikh, BLEURT: Learning robust metrics for text generation, in: ACL, 2020, pp. 7881–7892.

[75] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, CLIPScore: A reference-free evaluation metric for image captioning, in: EMNLP, 2021, pp. 7514–7528.

[76] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, et al., GLM-130B: An open bilingual pre-trained model, ICLR (2023).

[77] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: ICLR, 2019.

[78] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: ICML, 2021, pp. 8748–8763.

[79] Z. Li, B. Yang, Q. Liu, Z. Ma, S. Zhang, J. Yang, Y. Sun, Y. Liu, X. Bai, Monkey: Image resolution and text label are important things for large multi-modal models, in: CVPR, 2024, pp. 26763–26773.

[80] J.-X. Zhuang, X. Huang, Y. Yang, J. Chen, Y. Yu, W. Gao, G. Li, J. Chen, T. Zhang, Openmedia: Open-source medical image analysis toolbox and benchmark under heterogeneous ai computing platforms, in: PRCV, Springer, 2022, pp. 356–367.