A Benchmark for Multi-Task Evaluation of Pretrained Models in Medical Report Generation

Run Lin[1], Chunxiao Li[2*], Ruixuan Wang[1*]

[1] School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2] School of Foreign Languages, Guangdong Polytechnic Normal University

Abstract: MRG for medical images has become increasingly important due to the growing workload of radiologists in hospitals. However, current studies in the MRG field predominantly focus on specific modalities or training foundation models with a notable lack of research evaluating the impact of pre-trained models on performance across different tasks, particularly their cross-task capabilities. This study introduces a novel benchmark for medical multi-task learning that encompasses four medical modalities: CT, X-ray, ultrasound, and pathology. We believe this benchmark can provide a robust comparative basis for future research in this field. More importantly, we conduct an in-depth analysis comparing modality-specific pre-trained models, natural domain pre-trained models, and medical foundation pre-trained models. Our findings indicate that medical foundation pre-trained models generally outperform other pre-trained models across all tasks, while natural domain pre-trained models exhibit superior performance in cross-modality tasks. Our source code is available at https://github.com/Reckless0/MT-Med.git.

1 Introduction

Medical report generation (MRG) is a crucial application of image captioning technology [1] in the medical field. Designed to automatically produce accurate and coherent reports that detail the impressions and observations derived from medical images [2-4], effective MRG can significantly alleviate the diagnostic workload for physicians, decrease patient waiting time, and enhance the practical application of artificial intelligence in healthcare settings.

Recent advancements in vision and language applications have led to the rise of pre-trained models, which are designed for general applicability and demonstrate versatility across various tasks. However, in the MRG field, current studies primarily focus on a specific modality [5-9]. One issue is the paucity of research evaluating the impact of pre-trained models on performance across different tasks, particularly their cross-task capabilities despite the so many trained foundation modals proposed in some studies [10]. Another issue is that, despite the contributions of several open-source datasets to research in this field [11], there still lacks a multi-task dataset for systematic evaluation of the performance of models on different tasks.

To address the issues mentioned above, this study constructed a dataset for multiple radiological tasks to comprehensively evaluate the adaptability and robustness of models. Specifically, we leveraged the open-source ROCOv2 dataset [12] to extract three common imaging tasks: CT, X-Ray, and ultrasound, using modality keywords. In addition to that, the PatchGastricADC22 dataset [13] was incorporated as our fourth task. To ensure consistent data distribution across all tasks, we employed the GPT-4o API to augment captions for tasks with limited samples. Our proposed dataset is designed to support medical multi-task learning, incremental learning, and a wide range of application scenarios, providing a benchmark for research in medical multi-task analysis.

To comprehensively evaluate the performance of various MRG pre-trained models across multiple tasks, we utilized the Parameter-Efficient Fine-Tuning (PEFT) technique [14-15]. This approach significantly enhances model performance by fine-tuning a limited set of parameters while retaining the original model architecture and parameters. Consequently, it minimizes computational resource consumption and improves the model's generalization capabilities. Through an in-depth analysis of the experimental results across diverse tasks, we can identify the strengths and limitations of different MRG pre-trained models, thereby providing valuable insights for future research and applications.

The primary contributions of this paper are as follows:

1) We developed a publicly available benchmark for medical multi-task learning, which comprises images and captions for four medical modalities: CT, X-ray, ultrasound, and pathology. The dataset using GPT-4o was augmented to ensure consistent data distribution across all tasks.

2) We conducted experiments comparing medical modality-specific pre-trained models with their counterparts in the natural domain, as well as medical foundation pre-trained models. Using the PEFT technique, we fine-tuned

* Corresponding author: [a] wangruix5@mail.sysu.edu.cn
[b] lcx@gpnu.edu.cn

and evaluated these models on our proposed dataset. Valuable insights for both future research and practical applications are proposed therein.

2 Related works

In this section, we will review the related works on Medical Report Generation and Parameter-Efficient Fine-Tuning.

2.1 Medical Report Generation

In recent years, medical report generation has garnered increasing attention. To enhance model performance, researchers have pursued various improvements in different directions. Specifically, R2GenGPT [5] replaces the decoder of the traditional medical report generation framework with a more powerful large language model, thereby achieving superior performance. R2GenCSR [6] and XrayGPT [7] are recently proposed frameworks based on large language models (LLMs) for X-ray medical report generation (MRG). These frameworks employ Mamba as the visual backbone and retrieve contextual samples from the training set to enhance feature representation and discriminative learning. USFM [8] trains an ultrasound foundation model using a spatial-frequency dual masked image modeling method, which facilitates generalization across diverse ultrasound tasks. PLIP [9] leverages paired pathology images and captions from OpenPath through contrastive learning. BiomedCLIP and PMC-CLIP [10, 11], as multi-modal foundation models, incorporate domain-specific adaptations trained on their respective proposed datasets, thereby exhibiting superior performance across a broad range of downstream tasks.

2.2 Parameter-Efficient Fine-Tuning

With the proliferation of pre-trained foundation models [16, 17], efficiently adapting models to a specific task becomes a research hotspot. One effective technique is prompt engineering [18], which aims to affect the behaviors of language models by providing them with a textual template filled with task-related priors [19, 20], demonstrations of several examples [21, 22], or a chain of thoughts [23, 24]. 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 models. In implementation, PEFT either introduces lightweight components, e.g., Adapter, continuous prompts [14], and LoRA [15], vectors that scale the inner activations [25] into models, or adapts a small portion of inherent weights of models [26]. Recent practices utilizing these two techniques have demonstrated the effectiveness of adapting pre-trained models to the medical domain [27, 28].

3 Dataset Creation

3.1 Data Pre-processing

We acquired data from the open-source ROCOv2 dataset, which extracts image-caption pairs from the PMC Open Access Subset. This dataset comprises 79,789 radiological images accompanied by English captions. During the initial pre-processing, we employed SymSpell [29] to verify and correct the spelling of the English content, thereby enhancing caption quality. A filter was implemented to exclude sentences containing fewer than seven words, as shorter captions typically lack sufficient semantic information. Additionally, we filtered out excessive special symbols e.g. "@#&$%^". These steps ensure consistency and focus the evaluation on the linguistic content.

Following the initial preprocessing, the remaining data were subsequently categorized into three distinct radiological tasks: CT, X-ray, and ultrasound. Modality-specific keywords were utilized to determine the corresponding task for each image caption pair.

For the pathology task, we utilized the open-source dataset PatchGastricADC22[13], which contains approximately 260,000 pathology patch images, with an average of 260 images per caption, totaling around 1,000 captions. Given the substantial computational resources required to process all images simultaneously, we randomly selected 20 images per caption.

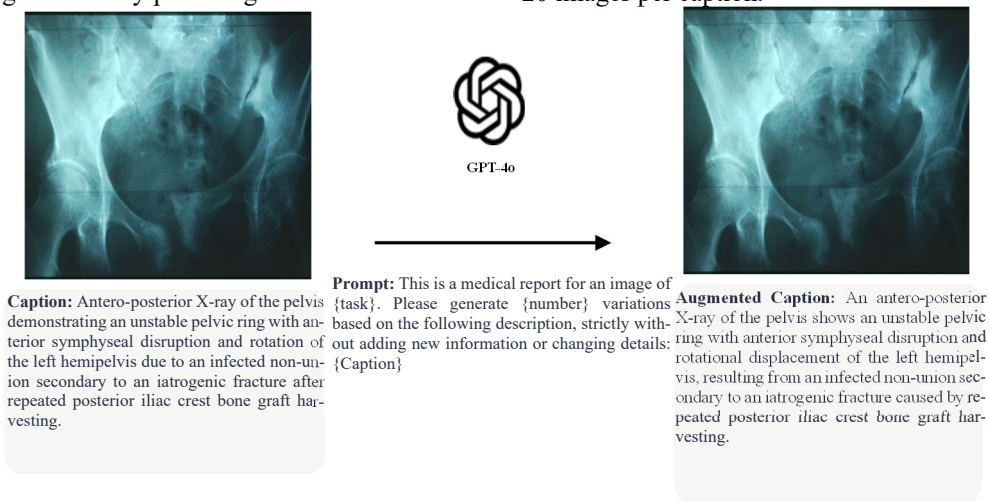

Figure 1. Caption Augmentation Workflow

3.2 GPT-4o Augmentation

To ensure the consistency of data distribution for each task, we utilized the GPT-4o API to augment captions. This approach not only maintains the consistent distribution of the dataset but also significantly enhances the generalization capability of the model. Specifically, by leveraging the GPT-4o API, we can generate new data samples that closely resemble the original caption in terms of semantic meaning, grammatical structure, and contextual logic. This ensures that the samples generated are highly consistent with the original data, thereby mitigating potential model bias caused by uneven data distribution. Furthermore, the robust language generation capabilities of the GPT-4o API introduce greater diversity into the dataset, enriching its content and providing a broader range of learning materials for the model, as shown in Figure 1.

3.3 Dataset Statistics

After the pre-processing steps, Table 1. presents statistics on our dataset. The sizes of the training, validation, and test sets across all tasks are within a similar range. Figure 2. shows the visualization of the number of tokens per caption in the training, validation, and test sets. The results reveal that almost all token counts fall within the range of 8 to 128 tokens, with a consistent distribution across all tasks. The consistent distribution ensures a balanced representation of each task throughout the training, validation, and test phases of the model evaluation process, thereby facilitating robust training and precise performance assessment.

Table 1. Task Image-caption Pairs Distribution

Task	Training-set	Validation-set	Test-set
CT	19301	3490	3444
X-Ray	14103	2285	2367
Ultrasound	15428	2866	2964
Pathology	12940	3220	3200

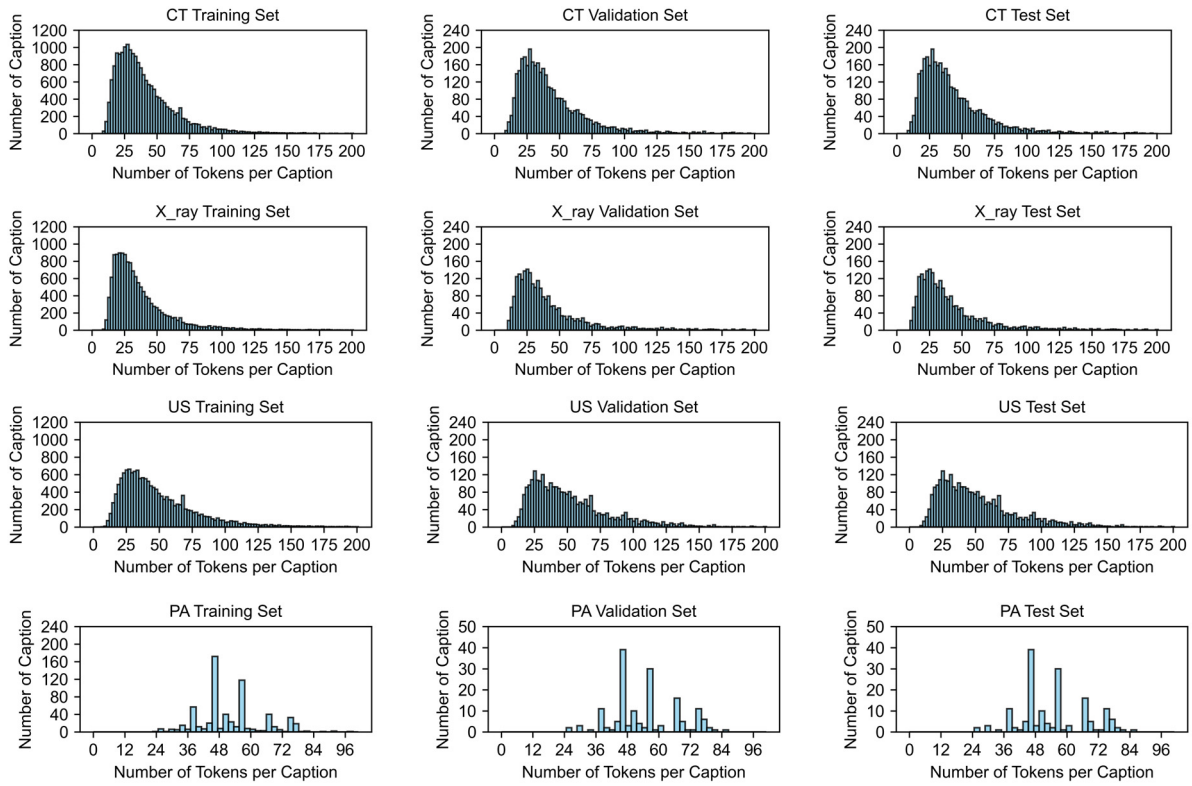

Figure 2. Histograms of the Number of Tokens per Caption (We use Llama's tokenizer [34] to split captions into tokens.)

4 Experiments

4.1 Experiment Setup

4.1.1 Metrics

We evaluated the model using BERTScore [30] and ROUGE-L [31], which are prominent metrics adopted in recent studies. Specifically, BERTScore is a model-based metric that quantifies the semantic similarity between two sentences. ROUGE-L measures the longest common subsequence between model-generated text and reference text. Additionally, we employed natural language generation (NLG) metrics, including BLEU, METEOR, and CIDEr, to comprehensively assess the performance of different models. In detail, BLEU evaluates text quality by

matching n-grams; CIDEr calculates the TF-IDF weights of n-grams in the generated and reference texts and compares them using cosine similarity; METEOR evaluates the generated text by aligning it with the reference and calculating a sentence-level similarity score. For all these metrics, higher scores indicate better performance. For statistical analysis, we evaluated 95% confidence intervals and calculated p-values for pairwise comparisons between the best model and other models each task.

4.1.2 Comparison with Pre-trained Models

We collected three types of open-source pre-trained models: (a) medical modality-specific models, (b) medical foundation models, and (c) natural domain pre-trained among models with the same architecture in (a) and (b). We selected SOTA or near SOTA models in their corresponding tasks, including XrayGPT, R2GenGPT, USFM, PLIP, BiomedCLIP. Table 2 presents the four experiments we conducted (Index #01 to #04) to compare pre-trained models of different medical modalities with their counterparts pre-trained on natural domains. Specifically, there are comparisons between Experiment #01 and the X-ray modality pre-trained model; Experiment #02 and the ultrasound modality pre-trained model; Experiment #03 and the pathology modality pre-trained model; Experiment #04 and the medical foundation pre-trained model. These comparisons were designed to investigate performance differences within the same modality (in-modality) and across different modalities (cross-modality).

4.1.3 Fine-tuning Settings

The pre-trained models above have been trained to develop a robust visual encoder, but the decoder varies. For instance, some models utilize a frozen Llama2-7B [32] as the language decoder, while others employ a BERT model. To ensure consistency and fairness, we adopted Llama2-7B as the language decoder. On this basis, we froze all parameters of the pre-trained model and applied standard fine-tuning methods to adapt these models to our dataset. Specifically, by leveraging PEFT technology, we added LoRA adapters into the attention layers of both the pre-trained encoder and decoder. This approach enables rapid adaptation to downstream tasks by adjusting a small subset of the model's inherent weights.

4.1.4 Hyper-Parameters

During fine-tuning, we utilized AdamW with an L2 weight decay of 0.05 to train the models using a batch size of 4 for 10 epochs. The learning rate was initialized at 1e-4 and followed a cosine annealing schedule. For evaluation, we employed the beam search decoding algorithm with a beam size of 3 to generate medical captions. To ensure the generated captions fall within a meaningful length range, we configured the decoder to produce at least 8 tokens and up to 128 tokens, as demonstrated by the histograms in Figure 2. We set the repetition penalty to 2.0 to mitigate redundancy and the length penalty to 2.0 to promote longer captions.

Table 2. Performance Metrics Obtained by Pre-trained Models on Different Tasks

Note: BS, R, B1, C and M denote BERTScore, ROUGE-L, BLEU-1, CIDEr and METEOR, respectively. 95% confidence intervals are included in brackets. The *italicized values* indicate statistically significant results ($p < 0.05$) based on comparisons between the BERTScore of the **Best model** and other models.

Index	Method	CT					X-Ray					Ultrasound					Pathology				
		BS	R	B1	C	M	BS	R	B1	C	M	BS	R	B1	C	M	BS	R	B1	C	M
#01	R2GenGPT	*0.622* (0.616, 0.628)	0.207 (0.198, 0.216)	0.251 (0.246, 0.257)	0.127 (0.122, 0.131)	0.097 (0.092, 0.102)	0.630 (0.626, 0.635)	0.195 (0.189, 0.201)	0.229 (0.221, 0.237)	0.166 (0.158, 0.173)	0.091 (0.087, 0.096)	0.621 (0.615, 0.627)	0.183 (0.176, 0.191)	0.201 (0.194, 0.208)	0.097 (0.088, 0.106)	0.077 (0.072, 0.083)	*0.819* (0.814, 0.825)	0.545 (0.537, 0.553)	0.610 (0.602, 0.619)	2.998 (2.948, 3.041)	0.346 (0.339, 0.353)
	Nature	*0.626* (0.622, 0.631)	0.215 (0.208, 0.221)	0.258 (0.252, 0.265)	0.147 (0.143, 0.151)	0.103 (0.099, 0.107)	*0.623* (0.617, 0.628)	0.194 (0.187, 0.200)	0.217 (0.209, 0.226)	0.147 (0.140, 0.155)	0.086 (0.081, 0.092)	0.625 (0.621, 0.630)	0.193 (0.187, 0.198)	0.218 (0.212, 0.225)	0.119 (0.111, 0.126)	0.083 (0.078, 0.088)	*0.825* (0.819, 0.830)	0.547 (0.540, 0.554)	0.618 (0.611, 0.625)	3.108 (3.065, 3.151)	0.350 (0.344, 0.357)
	XrayGPT	*0.623* (0.617, 0.629)	0.209 (0.202, 0.216)	0.246 (0.240, 0.251)	0.134 (0.128, 0.139)	0.096 (0.091, 0.100)	**0.632** (0.628, 0.635)	0.207 (0.204, 0.212)	**0.231** (0.224, 0.238)	0.174 (0.166, 0.181)	0.098 (0.094, 0.103)	0.621 (0.617, 0.627)	0.191 (0.185, 0.197)	0.207 (0.201, 0.214)	0.104 (0.096, 0.113)	0.075 (0.071, 0.080)	*0.815* (0.810, 0.821)	0.522 (0.513, 0.531)	0.611 (0.603, 0.619)	2.824 (2.758, 2.890)	0.339 (0.332, 0.346)
	Nature	*0.624* (0.619, 0.629)	0.212 (0.207, 0.218)	0.252 (0.247, 0.258)	0.142 (0.138, 0.147)	0.098 (0.094, 0.101)	*0.625* (0.621, 0.628)	0.201 (0.195, 0.207)	0.210 (0.201, 0.218)	0.152 (0.145, 0.160)	0.092 (0.086, 0.097)	0.622 (0.618, 0.628)	0.194 (0.188, 0.200)	0.216 (0.209, 0.222)	0.116 (0.109, 0.223)	0.081 (0.075, 0.087)	*0.822* (0.816, 0.828)	0.534 (0.526, 0.542)	0.617 (0.610, 0.625)	2.985 (2.926, 3.044)	0.346 (0.339, 0.350)
#02	USFM	*0.614* (0.610, 0.619)	0.197 (0.189, 0.204)	0.237 (0.232, 0.243)	0.114 (0.109, 0.118)	0.092 (0.089, 0.095)	*0.617* (0.614, 0.621)	0.185 (0.180, 0.191)	0.194 (0.187, 0.202)	0.121 (0.114, 0.127)	0.074 (0.068, 0.079)	0.624 (0.618, 0.629)	0.184 (0.177, 0.194)	0.213 (0.206, 0.221)	0.113 (0.107, 0.121)	0.082 (0.076, 0.087)	*0.806* (0.799, 0.813)	0.517 (0.508, 0.526)	0.581 (0.572, 0.591)	2.160 (2.096, 2.224)	0.302 (0.292, 0.312)
	Nature	*0.621* (0.615, 0.627)	0.204 (0.197, 0.212)	0.245 (0.239, 0.251)	0.125 (0.119, 0.130)	0.097 (0.093, 0.101)	*0.620* (0.615, 0.624)	0.187 (0.181, 0.192)	0.203 (0.195, 0.211)	0.132 (0.125, 0.138)	0.081 (0.076, 0.086)	0.621 (0.615, 0.626)	0.181 (0.174, 0.187)	0.194 (0.187, 0.201)	0.092 (0.081, 0.102)	0.075 (0.069, 0.081)	*0.812* (0.805, 0.820)	0.523 (0.514, 0.532)	0.588 (0.578, 0.597)	2.427 (2.332, 2.525)	0.314 (0.303, 0.324)
#03	PLIP	*0.619* (0.612, 0.625)	0.204 (0.198, 0.211)	0.226 (0.221, 0.232)	0.121 (0.116, 0.127)	0.094 (0.089, 0.099)	*0.621* (0.616, 0.627)	0.189 (0.183, 0.194)	0.210 (0.201, 0.219)	0.144 (0.136, 0.151)	0.095 (0.089, 0.101)	*0.618* (0.613, 0.624)	0.188 (0.181, 0.194)	0.210 (0.203, 0.218)	0.117 (0.109, 0.126)	0.082 (0.076, 0.089)	**0.839** (0.835, 0.844)	**0.599** (0.591, 0.607)	**0.643** (0.636, 0.649)	3.320 (3.287, 3.353)	**0.368** (0.361, 0.375)
	Nature	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
#04	Bio-MedCLIP	**0.632** (0.628, 0.637)	**0.223** (0.216, 0.230)	**0.271** (0.265, 0.274)	**0.205** (0.197, 0.213)	**0.117** (0.112, 0.123)	0.628 (0.625, 0.634)	**0.211** (0.206, 0.215)	0.227 (0.221, 0.234)	**0.174** (0.167, 0.181)	**0.104** (0.098, 0.111)	**0.625** (0.620, 0.629)	**0.198** (0.194, 0.203)	**0.234** (0.228, 0.239)	**0.133** (0.125, 0.141)	**0.093** (0.087, 0.099)	0.832 (0.826, 0.837)	0.593 (0.586, 0.601)	0.635 (0.629, 0.641)	**3.407** (3.374, 3.442)	0.358 (0.349, 0.366)
	Nature	0.625 (0.619, 0.632)	0.206 (0.198, 0.214)	0.232 (0.226, 0.239)	0.146 (0.139, 0.154)	0.096 (0.091, 0.112)	*0.623* (0.619, 0.628)	0.191 (0.184, 0.197)	0.217 (0.209, 0.225)	0.146 (0.138, 0.155)	0.098 (0.092, 0.105)	0.621 (0.617, 0.625)	0.190 (0.184, 0.195)	0.212 (0.205, 0.221)	0.115 (0.104, 0.126)	0.084 (0.075, 0.092)	*0.815* (0.808, 0.823)	0.533 (0.522, 0.543)	0.579 (0.565, 0.592)	2.529 (2.405, 2.653)	0.318 (0.311, 0.325)

Additionally, to enhance the decoder's ability to understand and generate content, we provided the language decoder with the text prompt: "Generate a comprehensive and detailed diagnosis report for this {task} image."

4.2 Results

Experiment results are presented in Table 2. In detail, we fine-tuned and tested pre-trained models on four tasks within our dataset: CT, X-ray, ultrasound, and pathology. For each task, we conducted the fine-tuning and testing experiments three times using random seeds 24, 42, and 2024, and report the average results in Table 2. The best and second-best results for each type of pre-trained model on the corresponding task are highlighted in **bold** and underlined respectively. Note that the natural modality model in Experiment #03 yields the same results as the one observed in Experiment #04 due to their same architecture.

5 Discussion

In all experimental tasks, the performance of the natural domain pre-trained models was comparable to that of the medical pre-trained models, and even slightly surpassed them in certain tasks. This suggests that despite the differences in visual features between medical and natural images, natural domain pre-trained models exhibit strong generalization capabilities in capturing generic features and can effectively adapt to the various tasks. This finding provides a novel perspective for medical report generation, indicating that reliance on medical pre-trained models may not always be necessary.

Comparing experiments #01, #02, and #03, we found that while medical modality-specific pre-trained models performed well in their respective modality's task, their performance in cross-modal tasks was inferior to that of natural domain pre-trained models. For instance, in experiment #03, the PLIP model demonstrated outstanding performance but showed limited effectiveness in cross-modal tasks e.g. CT and ultrasound tasks. This suggests that for tasks requiring multiple modalities, natural domain pre-trained models may better leverage their advantages, particularly in cross-modal tasks.

A plausible explanation for the superior performance of natural domain models in cross-modality tasks lies in their ability to generalize across diverse data types. Trained on extensive, heterogeneous datasets, these models acquire robust and transferable features that effectively bridge modality gaps. This strong generalization capability enables natural domain models to excel even when encountering new, unseen modalities, thereby enhancing their adaptability and versatility in cross-modality tasks.

By comparing experiments #01, #02, #03 with #04, we observed that medical foundation pre-training models generally outperformed modality-specific pre-training models across various tasks. This finding suggests that in the current field, using a foundation model pre-trained on large-scale medical datasets enhances overall performance more effectively than using a model pre-trained solely on specific modalities (e.g. X-ray or ultrasound). Foundation pre-training models are better at capturing universal features of medical images and cross-modal correlations, thereby improving performance across multiple medical tasks.

In Experiment #04, we observed that medical foundation pre-training models consistently outperformed natural domain pre-training models across all tasks. This result emphasizes the importance of prioritizing foundation models trained in multi-modal and multi-task environments over those limited to natural domains when selecting pre-training models. The superiority of medical foundation pre-training models stems from their capability to extract rich, multi-level features from diverse medical modalities, thereby exhibiting stronger cross-task and cross-modal generalization. In contrast, while natural domain models may excel in certain tasks, they generally exhibit weaker generalization and adaptability compared to medical foundation models. Consequently, researchers in the medical field should consider prioritizing medical foundation pre-training models to enhance the overall performance of medical report generation, particularly for diverse medical tasks.

Beyond technical performance, our framework significantly enhances the practicality of real-world MRG systems through two key aspects. First, the parameter-efficient fine-tuning (PEFT) strategy drastically lowers computational requirements for deployment. By reducing trainable weights by 95-99% compared to full fine-tuning, our framework enables cost-effective and resource-efficient deployment on standard clinical workstations. Second, our fully open-source benchmark leverages open-source data, providing a robust framework for privacy-aware deployment. While additional new modality or data become available in the future, leveraging PEFT technology will enable us to rapidly fine-tune pre-trained models for specific tasks without privacy issues.

6 Conclusion

In this study, we introduce a multi-task benchmark for medical report generation that encompasses CT, X-ray, ultrasound, and pathology tasks. To ensure consistent data distribution, we augmented the dataset using GPT-4o API. This benchmark aims to facilitate the evaluation of model performance across multiple tasks, thereby significantly advancing both academic research and technological development in the field. Additionally, we performed an in-depth analysis comparing modality-specific pre-trained models with natural domain pre-trained models, as well as medical foundation pre-trained models. Our findings provide several valuable insights. First, medical foundation pre-trained models almost outperformed modality-specific pre-trained models across all tasks. Second, while modality-specific pre-trained models excelled in their respective tasks, they exhibited inferior performance in cross-modal tasks compared to natural domain pre-trained models. Finally, our experiments demonstrated that medical foundation pre-trained models consistently surpassed natural domain pre-trained models across all tasks, underscoring the importance of prioritizing models trained in multi-modal

and multi-task environments. These models exhibit superior generalization capabilities across diverse medical tasks, making them more suitable for complex medical report generation applications.

In future work, we consider expanding our dataset to include additional tasks and developing an incremental model architecture to integrate knowledge from different tasks, further enhancing model performance.

References

1. Hossain M D Z, Sohel F, Shiratuddin M F, et al. (2019) A comprehensive survey of deep learning for image captioning[J]. ACM Computing Surveys, 51(6): 1-36.
2. Li, Y., Liang, X., Hu, Z., et al. (2018) Hybrid retrieval-generation reinforced agent for medical image report generation. Adv. Neural Inf. Process. Syst., 31.
3. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., et al. (2016) Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc., 23: 304–310.
4. Jing, B., Xie, P., Xing, E., et al. (2018) On the automatic generation of medical imaging reports. In: Association for Computational Linguistics, pp. 2577–2586.
5. Wang, Z., Liu, L., Wang, L., et al. (2023) R2GenGPT: Radiology report generation with frozen LLMs. Meta-Radiology, 1(3): 100033.
6. Wang, X., Li, Y., Wang, F., et al. (2024) R2GenCSR: Retrieving context samples for large language model-based X-ray medical report generation. arXiv preprint arXiv:2408.09743.
7. Thawkar, O., Shaker, A., Mullappilly, S.S., et al. (2023) XrayGPT: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971.
8. Jiao, J., Zhou, J., Li, X., et al. (2024) USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Med. Image Anal., 96: 103202.
9. Zuo, J., et al. (2023) PLIP: Language-image pre-training for person representation learning. arXiv preprint arXiv:2305.08386.
10. Zhang, S., et al. (2023) BiomedCLIP: A multimodal biomedical foundation model pre-trained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915.
11. Johnson, A.E.W., et al. (2019) MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data, 6: 317.
12. Rückert, J., et al. (2024) ROCOV2: Radiology objects in context version 2, an updated multimodal image dataset. Sci. Data, 11: 688.
13. Tsuneki, M., Kanavati, F. (2022) Inference of captions from histopathological patches. In: International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, pp. 1235–1250.
14. Liu, X., Ji, K., Fu, Y., et al. (2021) P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. arXiv preprint arXiv:2110.07602.
15. Hu, E.J., Shen, Y., Wallis, P., et al. (2021) LoRA: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685.
16. Zhao, W.X., Zhou, K., Li, J., et al. (2023) A survey of large language models. arXiv preprint arXiv:2303.18223.
17. Yang, W., Liu, M., Wang, Z., et al. (2024) Foundation models meet visualizations: Challenges and opportunities. Comput. Visual Media, 1–26.
18. Liu, P., Yuan, W., Fu, J., et al. (2023) Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, 55(9): 1–35.
19. Raffel, C., Shazeer, N., Roberts, A., et al. (2020) Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 21(140): 1–67.
20. Li, X.L., Liang, P. (2021) Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190.
21. Brown, T., Mann, B., Ryder, N., et al. (2020) Language models are few-shot learners. Adv. Neural Inf. Process. Syst., 33: 1877–1901.
22. Alayrac, J.B., Donahue, J., Luc, P., et al. (2022) Flamingo: A visual language model for few-shot learning. Adv. Neural Inf. Process. Syst., 35: 23716–23736.
23. Wei, J., Wang, X., Schuurmans, D., et al. (2022) Chain-of-thought prompting elicits reasoning in large language models. Adv. Neural Inf. Process. Syst., 35: 24824–24837.
24. Chowdhery, A., Narang, S., Devlin, J., et al. (2023) PaLM: Scaling language modeling with pathways. J. Mach. Learn. Res., 24(240): 1–113.
25. Liu, H., Tam, D., Muqeeth, M., et al. (2022) Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. Adv. Neural Inf. Process. Syst., 35: 1950–1965.
26. Zaken, E.B., Ravfogel, S., Goldberg, Y. (2021) BitFit: Simple parameter-efficient fine-tuning for transformer-based masked language models. arXiv preprint arXiv:2106.10199.
27. Singhal, K., Azizi, S., Tu, T., et al. (2023) Large language models encode clinical knowledge. Nature, 620(7972): 172–180.
28. Nori, H., King, N., McKinney, S.M., et al. (2023) Capabilities of GPT-4 on medical challenge problems. arXiv preprint arXiv:2303.13375.
29. Mon, E.P.P., Thu, Y.K., Yu, T.T., et al. (2021) SymSpell4Burmese: Symmetric delete spelling correction algorithm (SymSpell) for Burmese spelling

checking. In: 2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing. IEEE, pp. 1–6.

30. Zhang, T., Kishore, V., Wu, F., et al. (2019) BERTScore: Evaluating text generation with BERT. arXiv preprint arXiv:1904.09675.
31. Lin, C.Y. (2004) ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81.
32. Touvron H, Martin L, Stone K, et al. Llama 2: Open Touvron, H., et al. (2023) LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288.