DualPrompt-MedCap: A Dual-Prompt Enhanced Approach for Medical Image Captioning

Yining Zhao, Mukesh Prasad, and Ali Braytee

University of Technology Sydney, NSW 2007, Australia
Ali.Braytee@uts.edu.au

Abstract. Medical image captioning via vision-language models has shown promising potential for clinical diagnosis assistance. However, generating contextually relevant descriptions with accurate modality recognition remains challenging. We present DualPrompt-MedCap[1], a novel dual-prompt enhancement framework that augments Large Vision-Language Models (LVLMs) through two specialized components: (1) a modality-aware prompt derived from a semi-supervised classification model pre-trained on medical question-answer pairs, and (2) a question-guided prompt leveraging biomedical language model embeddings. To address the lack of captioning ground truth, we also propose an evaluation framework that jointly considers spatial-semantic relevance and medical narrative quality. Experiments on multiple medical datasets demonstrate that DualPrompt-MedCap outperforms the baseline BLIP-3 by achieving a 22% improvement in modality recognition accuracy while generating more comprehensive and question-aligned descriptions. Our method enables the generation of clinically accurate reports that can serve as medical experts' prior knowledge and automatic annotations for downstream vision-language tasks.

Keywords: Medical Image Captioning · Dual-prompt Enhancement · Semi-supervised Learning · Vision-Language Models

1 Introduction

Medical image understanding plays a crucial role in clinical diagnosis, yet the interpretation and reporting of such images still heavily rely on experienced pathologists [3], especially challenging in regions with limited expertise [16]. Vision-language models (VLMs) have shown promise in automating medical image captioning [11, 22], but face three critical limitations: (1) unreliable modality recognition, where models struggle to distinguish between imaging techniques [8]; (2) generic prompts lacking medical specificity; and (3) reliance on large annotated datasets, which are scarce due to privacy concerns [13].

Early captioning approaches adapted generic models like Show and Tell [21] to specialized architectures such as R2Gen [2], but typically overlooked modality awareness—crucial for diagnostic accuracy. While prompt-based methods [18]

[1] https://github.com/Yininnnnnng/DualPrompt-MedCap

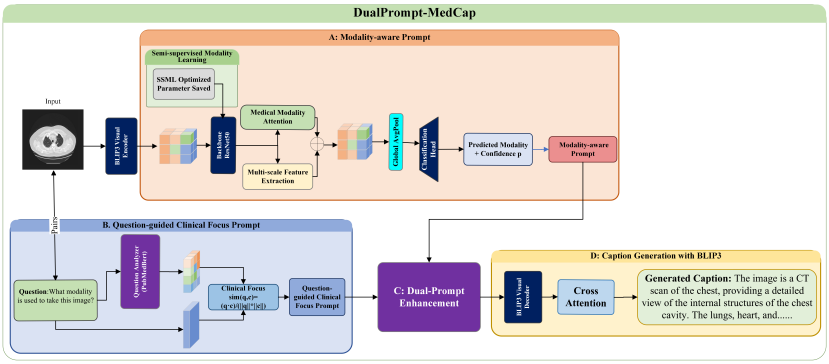

Fig. 1: Overview of the proposed DualPrompt-MedCap framework.

and BLIP-based models [10] show promise, they lack medical-specific constraints. The scarcity of labeled medical data compounds these challenges, with semi-supervised techniques like FixMatch [20] showing potential but remaining under-explored in medical contexts [14]. Additionally, evaluation metrics like BLEU [17] and domain-specific frameworks such as CheXpert [6] and RadGraph [7] rely on ground truth references and often fail to capture true clinical relevance.

To address these limitations, we propose DualPrompt-MedCap, a question-guided medical image captioning framework with three key innovations: (1) a semi-supervised learning approach with novel Medical Modality Attention mechanisms that significantly improves modality recognition accuracy with limited labeled data; (2) a dual-prompt strategy that integrates modality-aware prompts with question-guided clinical focus to generate contextually appropriate descriptions; and (3) a ground truth-independent evaluation framework that jointly assesses image relevance, question alignment, and adherence to radiological standards. Our experimental results demonstrate that this approach substantially outperforms baseline methods in generating clinically accurate and question-relevant medical image descriptions.

2 DualPrompt-MedCap Framework

We propose DualPrompt-MedCap with a parallel dual-pathway architecture: ResNet50 with Medical Modality Attention for modality classification and BLIP3 for visual understanding. These pathways merge only at prompt construction, addressing modality identification and question-guided description generation. Our framework (Fig. 1) leverages semi-supervised learning and specialized attention mechanisms, consisting of four main components: a modality-aware prompt generation module (A), a question-guided clinical focus analysis system (B), a dual-prompt enhancement mechanism (C), and a caption generation module powered by BLIP3 (D). By separating modality recognition from content

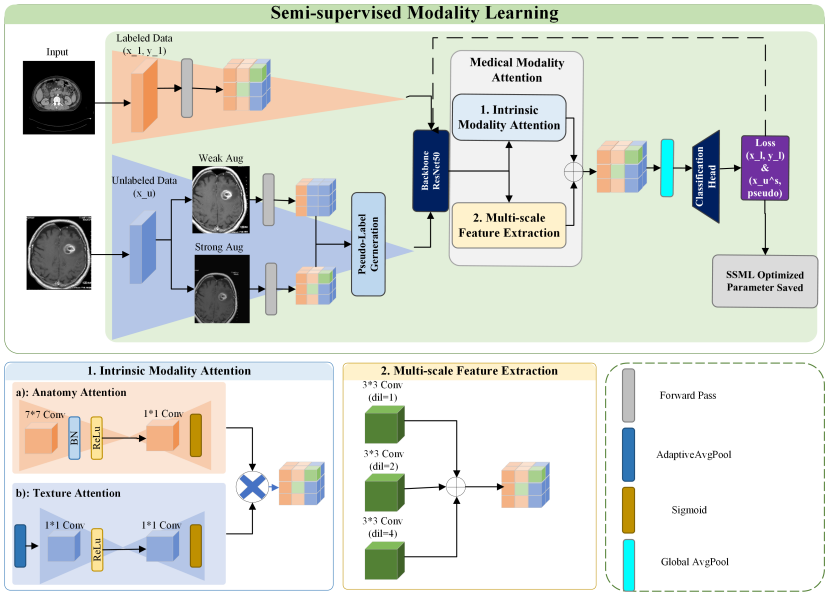

Fig. 2: Semi-supervised modality learning module.

description, DualPrompt-MedCap eliminates a common source of errors while focusing on diagnostically relevant features.

2.1 Modality-aware Prompt

The Modality-aware Prompt component leverages a pre-trained modality classifier for accurate imaging technique identification (CT, MRI, or X-ray). This classifier processes medical images through BLIP3's visual encoder followed by our enhanced backbone with Medical Modality Attention, incorporating the predicted modality into a structured prompt template.

Semi-supervised Modality Learning. We propose semi-supervised modality learning (Fig. 2) by extending FixMatch [20] to address limited labeled modality data. The framework processes labeled data (x_l, y_l) through a supervised pathway and unlabeled data (x_u) through weak-to-strong consistency training. For labeled data, we compute supervised loss using modality-specific class weights. For unlabeled data, weak augmentation generates pseudo-labels that guide learning from strongly augmented versions, incorporating only pseudo-labels exceeding a confidence threshold τ. Our implementation uses 202 labeled (75 MRI, 49 CT, 78 X-ray) and 2,042 unlabeled samples with class weights [1.5, 1.0, 1.0] to address MRI-CT confusion and τ=0.95 for optimal pseudo-label quality.

Medical Modality Attention. We introduce a Medical Modality Attention mechanism that adaptively emphasizes modality-specific visual cues through two components: Intrinsic Modality Attention and Multi-scale Feature Extraction. The feature transformation process is:

$$F_{\text{out}} = x \odot A_{\text{intr}}(x) + M_{\text{scale}}(x) \tag{1}$$

where A_{intr} represents Intrinsic Modality Attention, M_{scale} represents Multi-scale Feature Extraction, and $\odot$ denotes element-wise multiplication.

Intrinsic Modality Attention. The Intrinsic Modality Attention module captures anatomical structures and modality-specific imaging patterns:

$$A_{\text{intr}}(x) = A_{\text{anatomy}}(x) \odot A_{\text{texture}}(x) \tag{2}$$

a) Anatomy Attention: Implemented via 7×7 convolutions to enhance organ and tissue boundary detection, critical for distinguishing anatomical structures across modalities. Here, $A_{\text{anatomy}}(x)$ employs batch normalization and ReLU activation.

b) Texture Attention: Captures unique contrast patterns of each modality through channel-wise operations, where $A_{\text{texture}}(x)$ applies adaptive average pooling followed by two 1×1 convolutions with a dimension reduction and expansion strategy.

Multi-scale Feature Extraction. Processes features at different anatomical scales through dilated convolutions:

$$M_{\text{scale}}(x) = W_{\text{adj}} \left(\bigoplus_{d \in \{1,2,4\}} \text{Conv}_d(x) \right) \tag{3}$$

where Conv_d represents dilated convolutions with dilation rates of 1, 2, and 4, W_{adj} is 1×1 convolutions for channel adjustment, and $\bigoplus$ represents channel-wise concatenation.

Our semi-supervised strategy includes two key optimizations:

Modality-specific Augmentation: We address VLMs' tendency to misclassify MRI as CT through targeted augmentation strategies. For MRI, we apply enhanced geometric transformations and controlled intensity adjustments to preserve characteristic tissue contrast while increasing pattern diversity.

Semi-supervised Loss: We adapt FixMatch with a modified loss function for medical imaging:

$$\mathcal{L} = \sum_{(x_l, y_l)} w_{y_l} \text{CE}(f(x_l), y_l) + \lambda \sum_{x_u} \mathbb{I}(p > \tau) \cdot \text{CE}(f(x_u^s), \hat{y}_u) \tag{4}$$

where w_{y_l} are modality-specific class weights, $\hat{y}_u$ are pseudo-labels from weakly augmented samples, and τ (0.95) filters unreliable predictions.

Modality Prediction and Prompt Generation. During application, the framework loads pre-trained parameters to predict modality and construct modality-aware prompts.

2.2 Question-guided Clinical Focus Prompt

The Question-guided Clinical Focus Prompt component employs a Question Analyzer built on PubMedBERT [4] to extract semantic information from clinical queries. This component identifies both question types and relevant clinical concepts. The Question Analyzer computes clinical focus using cosine similarity between question embeddings and predefined clinical concept embeddings:

$$\text{sim}(q, c) = \frac{q \cdot c}{\|q\| \cdot \|c\|} \tag{5}$$

where q represents question embedding and c denotes embeddings of predefined clinical concepts. We construct concept dictionaries across six categories: anatomy (lung, heart, brain), pathology (tumor, nodule, lesion), location (left, right, upper), findings (normal, abnormal), measurements (mm, cm), and comparisons (increased, decreased). Each term is encoded using PubMedBERT to obtain 768-dimensional embeddings. Based on similarity scores, the analyzer generates contextual prompts such as "examining lung abnormalities" for pathology-focused queries. This process transforms semantic similarity into structured prompt text that guides BLIP3's generation.

The analyzer maintains comprehensive medical concept dictionaries covering anatomical structures, pathological findings, spatial locations, and comparison terms. Beyond semantic similarity, it conducts syntactic and contextual analysis to classify question types and extract relevant terms. For example, "Which side of the lung is abnormal?" is classified as a location-type question with anatomical and pathological components.

2.3 Dual-Prompt Enhancement

The Dual-Prompt Enhancement integrates the modality-aware prompt and question-guided clinical focus prompt into a unified guidance signal. This creates a comprehensive context capturing both imaging technique characteristics and clinical query focus. For instance, examining a chest CT with a question about lung abnormalities would create a prompt specifying both "CT scan of the chest" and "examining for lung abnormalities". By fusing these complementary prompts, our approach addresses the limitations of generic prompting methods, focusing caption generation on clinically relevant aspects and improving both accuracy and utility for medical professionals.

2.4 Caption Generation with BLIP3

The caption generation module powered by BLIP3 [23] transforms integrated dual prompts and visual features into coherent, clinically relevant descriptions.

BLIP3's visual decoder processes the input image's features, combined with dual-prompt information through cross-attention, enabling selective focus on image regions most relevant to both modality context and clinical question.

2.5 A Ground Truth-Independent Evaluation Framework

To address the limitations of reference-dependent metrics, we propose an automated evaluation framework that assesses both relevance and medical quality without requiring ground truth captions.

For relevance assessment, we use BiomedCLIP [24] to compute cosine similarities between image-caption ($S_{\text{image-text}}$) and question-caption ($S_{\text{question-text}}$) embeddings:

$$S_{\text{relevance}} = \alpha_1 \cdot S_{\text{image-text}} + \alpha_2 \cdot S_{\text{question-text}} \tag{6}$$

Medical quality evaluation encompasses three automated components:

$$S_{\text{quality}} = \beta_1 \cdot S_{\text{medical}} + \beta_2 \cdot S_{\text{clinical}} + \beta_3 \cdot S_{\text{structure}} \tag{7}$$

where: (1) S_{medical} measures UMLS [1] medical terminology usage via ScispaCy's [15] entity linker, automatically calculating entity density (entities/words ratio) and diversity (unique/total entities) through programmatic text processing without manual intervention; (2) S_{clinical} employs dictionary-based string matching to programmatically assess findings, anatomical localization, measurements, and comparisons with equal weights (25% each), ensuring objective evaluation without manual annotation; (3) $S_{\text{structure}}$ automatically evaluates report completeness through three equally-weighted aspects (33.3% each): explicit modality identification, multi-sentence description, and logical flow indicators like "suggest" or "indicate".

The final score combines both aspects:

$$S_{\text{final}} = \gamma_1 \cdot S_{\text{relevance}} + \gamma_2 \cdot S_{\text{quality}} \tag{8}$$

We set $\alpha_1 = \alpha_2 = 0.25$ for balanced relevance assessment, $\beta_1 = \beta_2 = \beta_3 = \frac{1}{3}$ for equal quality weighting, and $\gamma_1 = \gamma_2 = 0.5$ to ensure both relevance and quality contribute equally. This fully automated approach eliminates subjective manual evaluation while maintaining clinical validity [19, 13].

3 Experiments

All models were evaluated under identical zero-shot conditions without fine-tuning on the target datasets. We used consistent hyperparameters: beam size = 5, top_p = 0.9, and temperature = 1.0, to ensure fair comparison.

3.1 Datasets and Implementation Details

We evaluate on RAD [9] and SLAKE [12] datasets. RAD provides modality keywords for semi-supervised learning, while SLAKE offers complete ground truth for evaluation.

3.2 Experiment Settings

Our semi-supervised modality learning module uses ResNet50 initialized with ImageNet pre-trained weights, then trained on RAD dataset using our FixMatch-based approach. Learning rates of 1×10^{-5} (backbone) and 1×10^{-4} (attention modules) balance feature preservation with adaptation capabilities. For MRI images, we apply stronger geometric transformations ($\pm 15^{\circ}$ rotation, 15% translation) to address their unique characteristics and mitigate misclassification with CT scans. The FixMatch training strategy uses a confidence threshold of 0.95 for pseudo-labeling. For caption generation, we leverage BLIP3 as the base vision-language model and PubMedBERT for the Question Analyzer.

4 Results

Our experimental design validates each component's contribution through targeted evaluations: Table 1 demonstrates our modality recognition module's effectiveness achieving 98.30% accuracy (22% improvement over BLIP3), while Table 2 showcases the question-guided component's impact through superior question similarity (0.5694 vs 0.5323) and clinical accuracy (0.5833 vs 0.3348).

Table 1: Modality classification accuracy on SLAKE dataset.

Modality	**Semi-supervised**	**BLIP3 [23]**	**FixMatch [20]**	**Total Samples**
MRI	**92.11%**	11.84%	73.30%	472
CT	**100.00%**	92.37%	94.49%	361
X-Ray	**100.00%**	**100.00%**	**100.00%**	228
Average Accuracy	**98.30%**	77.66%	86.25%	1061

4.1 Modality Classification Results

We first demonstrate the ability of our proposed semi-supervised modality learning (FixMatch+Attention) to predict modality in datasets where this label does not exist, to be included in the prompt. Table 1 presents modality classification results on the SLAKE dataset. Our DualPrompt-MedCap framework significantly outperforms baseline methods, particularly in MRI recognition (92.11% vs BLIP3's 11.84%). The substantial performance gap (98.30% vs 77.66% average accuracy) confirms that specialized modality recognition is essential for generating clinically useful medical image captions.

Table 2: Evaluation results on SLAKE and RAD datasets.

Metric	SLAKE				RAD			
	Ours	**BLIP3**	**Tag2Text[5]**	**BLIP2 [10]**	**Ours**	**BLIP3**	**Tag2Text**	**BLIP2**
Final Score	**0.5396**	0.4716	0.2956	0.3273	**0.5337**	0.4494	0.3332	0.3318
Relevance Assessment								
Image Sim.	**0.3841**	0.3521	0.3070	0.3724	**0.3608**	0.3171	0.3249	0.3173
Question Sim.	**0.5694**	0.5323	0.5616	0.5585	**0.5929**	0.5013	0.5777	0.5728
Medical Quality Evaluation								
Med. Quality	0.3514	0.3545	**0.4593**	0.4476	0.3854	0.3717	0.4741	**0.4862**
Clin. Acc.	**0.5833**	0.3348	0.0272	0.0177	**0.5563**	0.3207	0.0114	0.0915
Structure	**0.8737**	0.8398	0.1482	0.1108	**0.8302**	0.8035	0.2492	0.1588

4.2 Medical Captioning Evaluation Results

Table 2 shows evaluation results on SLAKE and RAD datasets. DualPrompt-MedCap achieves the highest final scores on both datasets (0.5396 and 0.5337), outperforming all baseline models. The most significant improvements are in question similarity (0.5694/0.5929) and clinical accuracy (0.5833/0.5563), demonstrating our model's ability to generate captions aligned with clinical queries while maintaining medical relevance. Our approach also excels in structural consistency (0.8737/0.8302), in contrast to Tag2Text and BLIP2, which show poor clinical accuracy (<0.1) despite competitive medical terminology scores. These results validate that our dual-prompt approach significantly enhances clinical utility across different medical datasets.

4.3 Qualitative Analysis of DualPrompt-MedCap

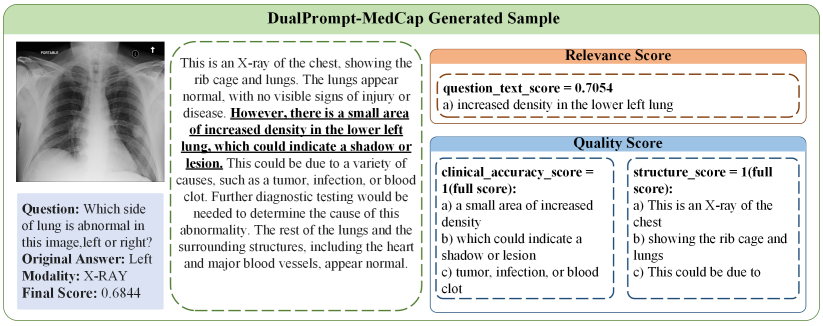

Fig. 3: DualPrompt-MedCap caption analysis showing relevance and quality.

Fig. 3 shows an analysis of a caption generated by DualPrompt-MedCap (final score: 0.6844). The caption directly addresses the query about lung abnor-

mality by identifying "increased density in the lower left lung," demonstrating strong question relevance (0.7054). The caption received perfect scores in clinical accuracy and report structure (1.0000 each), demonstrating how DualPrompt-MedCap integrates modality awareness and question guidance to generate both technically accurate and clinically meaningful descriptions.

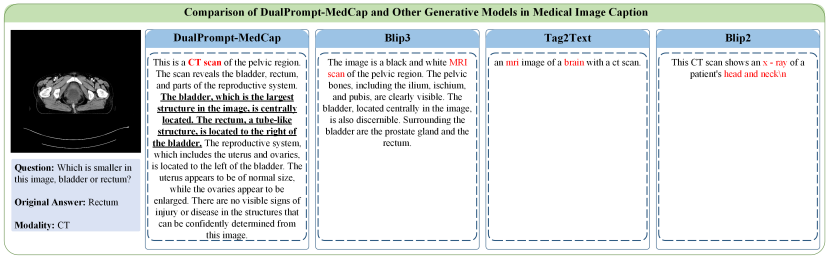

Fig. 4: DualPrompt-MedCap with other models in medical image captioning.

Fig. 4 compares captions from different models. While DualPrompt-MedCap accurately identifies the CT scan and addresses the clinical question, other models either misidentify the modality (BLIP3, Tag2Text, and BLIP2 all incorrectly label the imaging type) or fail to address the clinical query. DualPrompt-MedCap's caption includes precise anatomical descriptions, demonstrating how our dual-prompt strategy effectively integrates modality awareness and question relevance for clinically useful medical image descriptions.

5 Conclusion

In this paper, we proposed DualPrompt-MedCap, a framework for question-guided medical image captioning with enhanced modality recognition. By combining semi-supervised learning and a medical modality attention mechanism, we improved modality recognition accuracy. Our approach ensures both technical accuracy and clinical relevance, producing captions that meet radiological standards. Experimental results on VQA datasets highlight its superior performance, and potential to assist medical professionals, reduce radiologist workload, and improve diagnostic workflows, especially in resource-limited settings.

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

References

1. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)

2. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven transformer. arXiv preprint arXiv:2010.16056 (2020)
3. Farahani, N., Parwani, A.V., Pantanowitz, L.: Whole slide imaging in pathology: advantages, limitations, and emerging perspectives. Pathology and Laboratory Medicine International pp. 23–33 (2015)
4. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare (HEALTH) **3**(1), 1–23 (2021)
5. Huang, X., Zhang, Y., Ma, J., Tian, W., Feng, R., Zhang, Y., Li, Y., Guo, Y., Zhang, L.: Tag2text: Guiding vision-language model via image tagging. arXiv preprint arXiv:2303.05657 (2023)
6. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
7. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
8. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. arXiv preprint arXiv:1711.08195 (2017)
9. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 1–10 (2018)
10. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International conference on machine learning. pp. 19730–19742. PMLR (2023)
11. Li, Y., Liang, X., Hu, Z., Xing, E.P.: Hybrid retrieval-generation reinforced agent for medical image report generation. Advances in neural information processing systems **31** (2018)
12. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 1650–1654. IEEE (2021)
13. Liu, G., Hsu, T., McDermott, M., Boag, W., Weng, W., Szolovits, P., Ghassemi, M.: Clinically accurate chest x-ray report generation. corr. arXiv preprint arXiv:1904.02633 (2019)
14. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature biomedical engineering **5**(6), 555–570 (2021)
15. Neumann, M., King, D., Beltagy, I., Ammar, W.: Scispacy: fast and robust models for biomedical natural language processing. arXiv preprint arXiv:1902.07669 (2019)
16. Niazi, M.K.K., Parwani, A.V., Gurcan, M.N.: Digital pathology and artificial intelligence. The lancet oncology **20**(5), e253–e261 (2019)
17. 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)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)

19. Reiter, E., Belz, A.: An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Computational Linguistics **35**(4), 529–558 (2009)
20. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
21. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3156–3164 (2015)
22. Wang, X., Peng, Y., Lu, L., Lu, Z., Summers, R.M.: Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9049–9058 (2018)
23. Xue, L., Shu, M., Awadalla, A., Wang, J., Yan, A., Purushwalkam, S., Zhou, H., Prabhu, V., Dai, Y., Ryoo, M.S., et al.: xgen-mm (blip-3): A family of open large multimodal models. arXiv preprint arXiv:2408.08872 (2024)
24. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)