

UniCrossAdapter: Multimodal Adaptation of CLIP for Radiology Report Generation

Yaxiong Chen^{1,2}, Chuang Du^{1*}, Chunlei Li³, Jingliang Hu³, Yilei Shi³,
Shengwu Xiong^{1,2}, Xiao Xiang Zhu⁴, and Lichao Mou^{3(✉)}

¹ Wuhan University of Technology, Wuhan, China

² Shanghai Artificial Intelligence Laboratory, Shanghai, China

³ MedAI Technology (Wuxi) Co. Ltd., Wuxi, China
lichao.mou@medimagingai.com

⁴ Technical University of Munich, Munich, Germany

Abstract. Automated radiology report generation aims to expedite the tedious and error-prone reporting process for radiologists. While recent works have made progress, learning to align medical images and textual findings remains challenging due to the relative scarcity of labeled medical data. For example, datasets for this task are much smaller than those used for image captioning in computer vision. In this work, we propose to transfer representations from CLIP, a large-scale pre-trained vision-language model, to better capture cross-modal semantics between images and texts. However, directly applying CLIP is suboptimal due to the domain gap between natural images and radiology. To enable efficient adaptation, we introduce UniCrossAdapter, lightweight adapter modules that are incorporated into CLIP and fine-tuned on the target task while keeping base parameters fixed. The adapters are distributed across modalities and their interaction to enhance vision-language alignment. Experiments on two public datasets demonstrate the effectiveness of our approach, advancing state-of-the-art in radiology report generation. The proposed transfer learning framework provides a means of harnessing semantic knowledge from large-scale pre-trained models to tackle data-scarce medical vision-language tasks. Code is available at <https://github.com/chauncey-tow/MRG-CLIP>.

Keywords: report generation · CLIP · adapter.

1 Introduction

Radiology report writing is a tedious and error-prone task for radiologists due to the large volume of images needing interpretation. Automated report generation has recently emerged as a promising solution to expedite this process and alleviate the workload for radiologists. This task bears similarity to image captioning in computer vision, whereby textual descriptions must be produced to characterize visual inputs.

* Work done during an internship at MedAI Technology (Wuxi) Co. Ltd.

There has been growing interest in this direction. The authors of [1] propose to generate radiology reports with a memory-driven Transformer and firstly conduct studies on MIMIC-CXR dataset [2]. They later augment their model with a cross-modal memory module [3]. [4] puts forth an approach to distill both posterior and prior knowledge to further boost performance. In order to better align visual and textual features, [5] employs reinforcement learning over the cross-modal memory network [3]. In [6], the authors design a cross-modal prototype network to facilitate interactions across modalities. Aiming to promote semantic alignment, [7] explicitly leverage text embeddings to guide visual feature learning. Recently, [8] introduces a framework that makes use of a dynamic graph to enhance visual representations in a contrastive learning paradigm for radiology report generation tasks.

Due to medical privacy concerns, the difficulty of gathering medical data, and the labor-intensive nature of annotation, the amount of data available for radiology report generation is relatively small compared to that used for image captioning in computer vision. For example, IU-Xray (4K images) [9] and MIMIC-CXR (220K images) [2] are much smaller than image captioning datasets Conceptual Captions (3.3M images) [10] and Conceptual 12M (12M images) [11]. Learning comprehensively from such limited data makes it challenging for current methods to fully understand cross-modal semantics between radiological images and reports [1, 3–8]. Overcoming this paucity of labeled data to better learn these semantics is crucial for advancing radiology report generation.

Recently, leveraging large-scale pre-trained vision-language models, such as CLIP [12], which is trained on 400 million image-text pairs collected from the internet to match images with their corresponding textual descriptions, has become a promising approach for tackling downstream tasks in computer vision. However, the application of such models on radiology report generation still remains unexplored. In this work, we propose transferring the knowledge encapsulated in CLIP to the task of automatic report generation to better model the semantic relationship between medical images and their associated radiological findings.

Despite its strong performance, directly applying CLIP to radiology report generation tasks poses certain challenges. CLIP has been pre-trained on large-scale natural image-text datasets, exhibiting a substantial domain divergence from medical images. Therefore, while the model encapsulates rich semantic knowledge about everyday scenes, fine-tuning is imperative to adapt CLIP to radiology. However, conducting a full fine-tuning of a model as massive as CLIP is highly impractical given immense computational demands. To enable efficient adaptation, we propose uni- and cross-modal adapter (UniCrossAdapter), a parameter-efficient fine-tuning approach to adapt CLIP for the task of radiology report generation. The key idea is to integrate lightweight adapter modules into CLIP that can be fine-tuned on the target task while keeping the pre-trained backbone parameters frozen. The modules are distributed to both visual and textual modalities and their interactions for better aligning medical images and texts. Our contributions are three-fold.

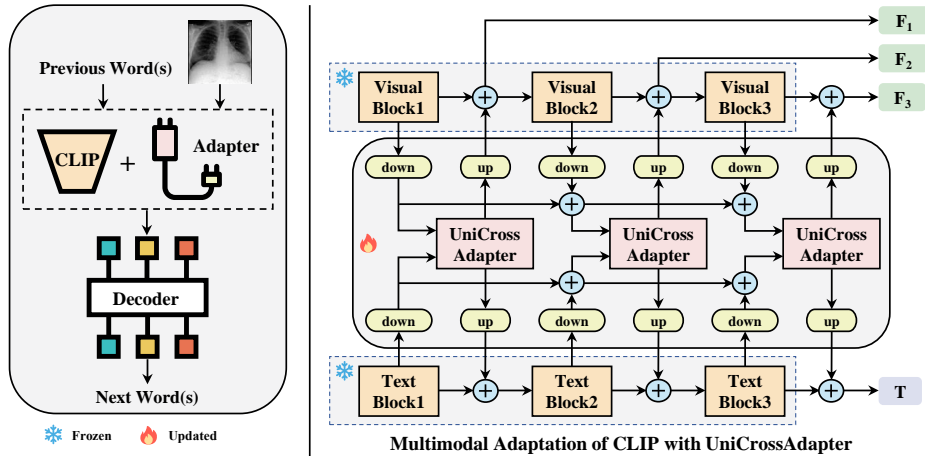


Fig. 1. (Left) Overall architecture of our method for radiology report generation, leveraging CLIP and the proposed UniCrossAdapter. (Right) Illustration of the interaction between the UniCrossAdapter and CLIP’s text and image encoders.

- We investigate the transfer of representations learned by CLIP to describe medical image findings.
- We introduce a novel adapter architecture that improves vision-language alignment on radiology images and reports by coupling image and text adapter modules through a cross-attention mechanism.
- Our approach achieves state-of-the-art performance on IU-Xray and MIMIC-CXR, the two most used benchmark datasets.

2 Method

We propose an end-to-end framework for automatic radiology report generation, as illustrated in Fig. 1. The model comprises two key components: (i) the adaptation of CLIP with UniCrossAdapter to learn visual and textual representations for radiology data, and (ii) a decoder that generates reports. In what follows, we first detail each of them. Then, we describe the training and inference procedures.

2.1 Multimodal Adaptation of CLIP with UniCrossAdapter

Recent work has explored parameter-efficient fine-tuning methods [13–19] for adapting large pre-trained models to downstream tasks. However, architectures used in prior efficient tuning techniques, e.g., down-up feedforward layers [18, 19] and LoRA [17], may be too simple to effectively adapt complex multimodal models. Moreover, most existing approaches have focused largely on unimodal

or basic classification tasks, with little exploration on more challenging multi-modal setups requiring inter-modality interaction modeling. Our proposed Uni-CrossAdapter is dedicated to the multimodal adaptation of CLIP.

CLIP’s Text and Image Encoders We utilize the pre-trained CLIP text Transformer to extract text features. Due to its large parameter size, the text Transformer remains frozen during fine-tuning. We then evenly split it into three sequential blocks and denote the text feature map from each block as $\mathbf{T}_i \in \mathbb{R}^{N \times D}$, where $i \in \{1, 2, 3\}$, N is the number of tokens, and D is the feature dimension.

For the visual branch, we use CLIP’s image encoder, specifically ResNet-101, to extract multi-scale visual features \mathbf{F}_i from the last three stages. Similar to the text encoder, we freeze the image encoder during fine-tuning to leverage rich semantics learned from pre-training.

Unimodal and Cross-Modal Adaptation The visual and linguistic features are first projected to a lower-dimensional space. Residual connections are further formed between consecutive adapter layers to enrich unimodal representations. This process can be formulated as

$$\begin{aligned}\hat{\mathbf{F}}_i &= \text{down}(\mathbf{F}_i) + \hat{\mathbf{F}}_{i-1}, \\ \hat{\mathbf{T}}_i &= \text{down}(\mathbf{T}_i) + \hat{\mathbf{T}}_{i-1},\end{aligned}\tag{1}$$

where $\text{down}(\cdot)$ indicates dimension reduction layers implemented by convolutional and linear layers for visual and textual features, respectively. To encourage interactions within each modality, we apply multi-head self-attention (MHSA) on both modalities:

$$\begin{aligned}\mathbf{F}_i^{sa} &= \text{MHSA}(\hat{\mathbf{F}}_i), \\ \mathbf{T}_i^{sa} &= \text{MHSA}(\hat{\mathbf{T}}_i).\end{aligned}\tag{2}$$

For coupling the visual and linguistic adapter modules, we perform multi-head cross-attention (MHCA) across the adapted unimodal representations for establishing cross-modal interactions:

$$\begin{aligned}\mathbf{F}_i^{ca} &= \text{FFN}(\text{MHCA}(Q = \mathbf{F}_i^{sa}, K = \hat{\mathbf{T}}_i, V = \hat{\mathbf{T}}_i)), \\ \mathbf{T}_i^{ca} &= \text{FFN}(\text{MHCA}(Q = \mathbf{T}_i^{sa}, K = \hat{\mathbf{F}}_i, V = \hat{\mathbf{F}}_i)).\end{aligned}\tag{3}$$

Then, we incorporate the interacted features into the original features:

$$\begin{aligned}\tilde{\mathbf{F}}_i &= \text{up}(\mathbf{F}_i^{ca}) + \mathbf{F}_i, \\ \tilde{\mathbf{T}}_i &= \text{up}(\mathbf{T}_i^{ca}) + \mathbf{T}_i,\end{aligned}\tag{4}$$

where $\text{up}(\cdot)$ denotes dimension recovery implemented by deconvolution and linear layers.

Feature Modulation and Multi-Scale Fusion Since radiology images contain multi-scale anatomical structures (e.g., lung and heart) that require model attention, we fuse the multi-scale visual features to obtain comprehensive representations. Before fusion, we modulate the visual features of different scales by interacting a global text feature τ , obtained via a projection layer in the text Transformer, with each \tilde{F}_i to highlight relevant regions:

$$\begin{aligned} M_i &= \text{MHCA}(Q = s(\tilde{F}_i), K = \tau, V = \tau), \\ Z &= \text{Conv}_{1 \times 1} \circ \text{Concat}(M_1, M_2, M_3), \end{aligned} \quad (5)$$

where s denotes a convolutional layer to project the multi-scale features to a unified scale. M_i represents the modulated visual features. \circ is a composition function, and $Z \in \mathbb{R}^{C \times H \times W}$ is the fused visual feature.

In addition, to incorporate spatial information into Z , we concatenate it with spatial coordinates $P \in \mathbb{R}^{2 \times H \times W}$ across the channel dimension. The resulting feature is then passed through a 3×3 convolutional layer to reduce the enlarged channel dimension. This process can be written as

$$X = \text{Conv}_{3 \times 3} \circ \text{Concat}(Z, P). \quad (6)$$

Finally, we send X into a vision Transformer [20] network such that X is transformed to a sequence of feature vectors $\{v_1, v_2, \dots, v_N\}$, where $v_i \in \mathbb{R}^D$ for the following procedure.

2.2 Report Decoder

We adopt a standard Transformer decoder [21] to generate reports. The decoder takes as input the adapted, fused multimodal representations from the CLIP-driven image and text encoders, and generates tokens autoregressively.

2.3 Training and Inference

Training Let I be an input radiology image, and its ground truth report is denoted as $R = \{[\text{SOS}], w_1, w_2, \dots, w_L, [\text{EOS}]\}$, where $w_i \in \mathcal{V}$ represents the i -th token and \mathcal{V} is the vocabulary set. $[\text{SOS}]$ and $[\text{EOS}]$ are the appended start and end tokens, while L is the length of the sequence. At training time, we first feed I and $\{[\text{SOS}], w_1, w_2, \dots, w_L\}$ into the image and text encoders with our adapter to derive a multimodal representation. The Transformer decoder then takes the multimodal representation as input and $\{[\text{SOS}], w_1, w_2, \dots, w_L\}$ as query to generate a predicted token sequence $\{p_1, p_2, \dots, p_L, p_{L+1}\}$. We optimize the model by minimizing the cross entropy loss between the predicted sequence and the corresponding ground truth sequence $\{w_1, w_2, \dots, w_L, [\text{EOS}]\}$:

$$\mathcal{L}_{\text{ce}} = -\frac{1}{L+1} \sum_{i=1}^{L+1} w_i \log(p_i). \quad (7)$$

Table 1. Comparison results on the IU-Xray and MIMIC-CXR datasets. * denotes results replicated from official code. † indicates replicated results without pre-training on the datasets. **Bold** indicates the best results, and underline indicates the second best results.

Dataset	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
IU-Xray	R2Gen	0.470	0.304	0.219	0.165	0.371	0.187
	SentSAT+KG	0.441	0.291	0.203	0.147	0.367	-
	CMCL	0.473	0.305	0.217	0.162	0.378	0.186
	\mathcal{M}^2 Tr. Prog.	0.486	0.317	0.232	0.173	0.390	0.192
	CMN	0.475	0.309	0.222	0.170	0.375	0.191
	PPKED	0.483	0.315	0.224	0.168	0.376	-
	CMM+RL	0.494	0.321	<u>0.235</u>	<u>0.181</u>	0.384	0.201
	XPRONET*	0.491	<u>0.325</u>	0.228	0.169	0.387	<u>0.202</u>
	DCL	-	-	-	0.163	0.383	0.193
	M2KT	<u>0.497</u>	0.319	0.230	0.174	0.399	-
	VLCI†	0.324	0.211	0.151	0.115	0.379	0.166
	RAMT	0.482	0.310	0.221	0.165	0.377	0.195
	PromptMRG	0.401	-	-	0.098	0.281	0.160
	Ours	0.509	0.349	0.257	0.195	<u>0.395</u>	0.210
MIMIC-CXR	R2Gen	0.353	0.218	0.145	0.103	0.277	0.142
	CMCL	0.344	0.217	0.140	0.097	0.281	0.133
	\mathcal{M}^2 Tr. Prog.	0.378	0.232	0.154	0.107	0.272	0.145
	CMN	0.353	0.218	0.148	0.106	0.278	0.142
	PPKED	0.360	0.224	0.149	0.106	0.284	0.149
	CMM+RL	0.381	0.232	0.155	0.109	<u>0.287</u>	0.151
	XPRONET	0.344	0.215	0.146	0.105	0.279	0.138
	DCL	-	-	-	0.109	0.284	0.150
	M2KT	<u>0.386</u>	0.237	<u>0.157</u>	0.111	0.274	-
	VLCI†	0.357	0.216	0.144	0.103	0.256	0.136
	RAMT	0.362	0.229	0.157	<u>0.113</u>	0.284	<u>0.153</u>
	PromptMRG	0.398	-	-	0.112	0.268	0.157
	Ours	0.375	0.237	0.165	0.120	0.289	0.134

Inference During inference, our model generates texts in an autoregressive manner. Given a test image, the model is first provided an [SOS] token as a prompt to predict the first token. The predicted first token is then concatenated with the [SOS] token as a new prompt to predict the second token. This process continues iteratively, with the previously predicted token(s) and [SOS] token as a prompt to predict each subsequent token, until an [EOS] token is predicted indicating the end of generation. This autoregressive way allows the model to condition each token prediction on its previous predictions, yielding more coherent and fluent text.

3 Experiments

3.1 Datasets and Evaluation Metrics

We conduct experiments on two datasets: IU-Xray [9] and MIMIC-CXR [2]. IU-Xray comprises 7,470 chest X-ray images along with 3,955 radiology reports.

We tokenize words with > 3 occurrences and truncate/pad reports to 60 tokens. MIMIC-CXR is a large-scale chest X-ray dataset containing 473,057 radiographs with 206,563 associated reports. Tokens with frequency > 10 are retained, and reports are truncated/padded to 78 tokens to conform with CLIP’s specifications. For a fair and consistent evaluation on the two datasets, we use the same data splits as employed in prior works [1, 3–8, 22, 23].

We evaluate report generation quality using standard natural language processing metrics: BLEU 1-4, METEOR, and ROUGE-L. All metrics are computed with a standard evaluation toolkit [24].

3.2 Implementation Details

The MHSAs and MHCAs in UniCrossAdapter use 64-dim features and 4 attention heads. For IU-Xray, the vision Transformer and report decoder have 3 layers each, while for MIMIC-CXR, we use 6 layers due to its larger size. To mitigate IU-Xray’s limited data, we use a consolidated vocabulary combining both datasets, enabling more diverse word projections. We choose Adam as the optimizer and use a batch size of 16 for training. We employ an initial learning rate of $1e-5$ and weight decays of $5e-5$ and $4e-5$ for IU-Xray and MIMIC-CXR, respectively. We also apply dropout for regularization with rates of 0.09 and 0.1 for the IU-Xray and MIMIC-CXR datasets, respectively.

3.3 Comparison with State-of-the-Art Methods

We compare against existing methods including R2Gen [1], SentSAT+KG [25], CMCL [26], \mathcal{M}^2 Tr. PROGRESSIVE [27], CMN [3], PPKE [4], CMM+RL [5], XPRONET [6], DCL [8], M2KT [7], VLCT [28], RAMT [29], and PromptMRG [30]. As shown in Table 1, the proposed approach outperforms the best competing method by 2.4% in BLEU-2, 2.2% in BLEU-3, 1.4% in BLEU-4, 1.2% in BLEU-1, and 0.8% in METEOR on IU-Xray. While slightly lower in ROUGE-L compared to M2KT [7], our method remains the top performer overall. On the larger MIMIC-CXR dataset, our model also shows improvements of 0.8% in BLEU-3 and 0.7% in BLEU-4 compared to prior art, along with comparable BLEU-2 and ROUGE-L. As evidenced in previous work [1, 3–8, 25–30], gains on MIMIC-CXR are more marginal due to its scale. Overall, our approach achieves state-of-the-art or comparable performance on both IU-Xray and MIMIC-CXR datasets.

3.4 Ablation Study

We ablate key components of our model, UniCrossAdapter and CLIP encoders, to analyze their impact quantitatively (cf. Table 2). Removing either significantly degrades performance, validating their efficacy. This suggests that CLIP’s multimodal knowledge facilitates learning cross-modal semantic alignments.

Fig. 2 shows example radiology reports generated by our full model and its ablated versions. In the absence of either UniCrossAdapter or CLIP pre-training

Table 2. Ablation results on the IU-Xray and MIMIC-CXR datasets. The best results are in **bold**. w/o denotes “without”.

IU-Xray	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
w/o UniCrossAdapter	0.302	0.201	0.146	0.109	0.375	0.154
w/o CLIP pre-training weights	0.450	0.298	0.208	0.147	0.357	0.188
Full model	0.509	0.349	0.257	0.195	0.395	0.210
MIMIC-CXR	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-L	METEOR
w/o UniCrossAdapter	0.087	0.055	0.038	0.028	0.226	0.077
w/o CLIP pre-training weights	0.351	0.196	0.118	0.077	0.250	0.118
Full model	0.375	0.237	0.165	0.120	0.289	0.134

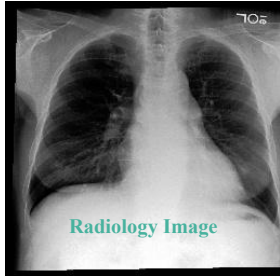
	Generation Results	
Ground Truth the cardiomeastinal and hilar contours are normal. the lungs are well expanded and clear without focal consolidation pleural effusion or pneumothorax. mild degenerative changes are seen in the thoracic spine.	Ours w/o UniCrossAdapter the lungs are clear. there is no pleural effusion or pneumothorax. the lungs are clear. the lungs are. there is no pneumothorax. the lungs are. the right are.	
	Ours w/o CLIP pre-training weights no change. the heart is normal. no the heart is normal. the lungs are clear. no pleural effusion or pneumothorax. the heart size is normal. the mediastinal and hilar contours are normal. no acute osseous abnormalities. no acute osseous abnormalities.	
	Ours pa and lateral views of the chest were obtained. the lungs are clear. there is no focal consolidation pleural effusion or pneumothorax. the cardiomeastinal and hilar contours are unremarkable. there is no pulmonary edema. the cardiomeastinal silhouette is normal. there is no acute osseous abnormalities.	

Fig. 2. Example of radiology report generation results on a test image from our model and its ablated variants. Ground truth words present in the generated reports are highlighted in color.

weights, the model produces noticeably inferior results. Specifically, the generated results exhibit poor grammar and a high level of repetition. This demonstrates that introducing pre-trained cross-modal knowledge from CLIP into the task of radiology report generation proves highly effective for producing more comprehensive and fluent reports. Moreover, this also highlights the significance of vision-language alignment by our adapter method for the overall model. Furthermore, we observe that reports generated by our full model demonstrate a level of professionalism comparable to ground truths.

4 Conclusion

In this work, we propose leveraging CLIP for the task of automated radiology report generation. Recognizing the infeasibility of fully fine-tuning such a

massive model, we introduce UniCrossAdapter, a parameter-efficient fine-tuning approach to adapt CLIP to this domain. Our experiments demonstrate state-of-the-art performance on two public benchmarks. Qualitative analysis shows our model is capable of generating coherent reports describing key clinical findings in medical images. This work illustrates the promise of large pre-trained multi-modal models for radiology report generation and introduces a method to make their adoption practical.

Acknowledgments. This work is supported in part by the National Key Research and Development Program of China (2022ZD0160604), in part by the Natural Science Foundation of China (62101393/62176194), in part by the High-Performance Computing Platform of YZBSTCACC, and in part by MindSpore (<https://www.mindspore.cn>), a new deep learning framework.

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

References

1. Chen, Z., Song, Y., Chang, T.H., Wan, X.: Generating radiology reports via memory-driven Transformer. In: 2020 Conference on Empirical Methods in Natural Language Processing, pp. 1439–1449. (2020)
2. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
3. Chen, Z., Shen, Y., Song, Y., Wan, X.: Cross-modal memory networks for radiology report generation. In: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 5904–5914. (2022)
4. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13753–13762. (2021)
5. Qin, H., Song, Y.: Reinforced cross-modal alignment for radiology report generation. In: Findings of the Association for Computational Linguistics: ACL 2022, pp. 448–458. (2022)
6. Wang, J., Bhalerao, A., He, Y.: Cross-modal prototype driven network for radiology report generation. In: European Conference on Computer Vision, pp. 563–579. (2022)
7. Yang, S., Wu, X., Ge, S., Zheng, Z., Zhou, S.K., Xiao, L.: Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis* **86**, 102798 (2023)
8. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest X-ray report generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3334–3343. (2023)
9. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association* **23**(2), 304–310 (2016)

10. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: The 56th Annual Meeting of the Association for Computational Linguistics, pp. 2556–2565. (2018)
11. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3558–3568. (2021)
12. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. (2021)
13. Guo, D., Rush, A.M., Kim, Y.: Parameter-efficient transfer learning with diff pruning. In: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 4884–4896. (2021)
14. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* **132**(2), 581–595 (2024)
15. Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J., Luo, P.: AdaptFormer: Adapting vision Transformers for scalable visual recognition. In: Advances in Neural Information Processing Systems, pp. 16664–16678. (2022)
16. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. *International Journal of Computer Vision* **130**(9), 2337–2348 (2022)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021)
18. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: International Conference on Machine Learning, pp. 2790–2799. (2019)
19. Rücklé, A., Geigle, G., Glockner, M., Beck, T., Pfeiffer, J., Reimers, N., Gurevych, I.: AdapterDrop: On the efficiency of adapters in Transformers. In: 2021 Conference on Empirical Methods in Natural Language Processing, pp. 7930–7946. (2021)
20. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 6000–6010. (2017)
22. Kong, M., Huang, Z., Kuang, K., Zhu, Q., Wu, F.: TranSQ: Transformer-based semantic query for medical report generation. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 610–620. (2022)
23. Li, J., Li, S., Hu, Y., Tao, H.: A self-guided framework for radiology report generation. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 588–598. (2022)
24. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015)
25. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: AAAI Conference on Artificial Intelligence, pp. 12910–12917. (2020)

26. Liu, F., Ge, S., Wu, X.: Competence-based multimodal curriculum learning for medical report generation. In: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 3001–3012. (2021)
27. Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive Transformer-based generation of radiology reports. In: Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 2824–2832. (2021)
28. Chen, W., Liu, Y., Wang, C., Li, G., Zhu, J., Lin, L.: Visual-linguistic causal intervention for radiology report generation. arXiv preprint arXiv:2303.09117 (2023)
29. Zhang, K., Jiang, H., Zhang, J., Huang, Q., Fan, J., Yu, J., Han, W.: Semi-supervised medical report generation via graph-guided hybrid feature consistency. *IEEE Transactions on Multimedia* **26**, 904–915 (2024)
30. Jin, H., Che, H., Lin, Y., Chen, H.: PromptMRG: Diagnosis-driven prompts for medical report generation. In: AAAI Conference on Artificial Intelligence, pp. 2607–2615. (2024)