Knowledge Graph-enhanced Vision-to-Language Multimodal Models for Radiology Report Generation

Yongli Mou[1][0000−0002−2064−0107]

RWTH Aachen University, Ahornstr. 55, 52074 Aachen, Germany
mou@dbis.rwth-aachen.de

Abstract. Current deep learning models for automated radiology report generation leverage architecture that comprises a visual encoder and a text decoder, but often lack the semantic depth and contextual understanding necessary for producing clinically relevant, easy-to-read, and accurate reports. The situation is even more challenging due to the complex nature of medical imaging and the specialized language and medical terminologies in radiology reports. The gap in domain-specific knowledge in current deep learning models underscores the necessity for approaches that integrate specialized radiological expertise into advanced language models. In this research, we propose a knowledge graph-enhanced vision-to-language multimodal model for radiology report generation, that leverages existing medical and radiological knowledge graphs. We explore contrastive learning approaches for pre-training multimodal models to learn the joint embeddings of modalities including images, graphs and texts. Our research not only contributes to the field of semantic web research by demonstrating the potential of knowledge graphs in enhancing deep learning models but also aims to revolutionize the radiology reporting process by automating it with greater accuracy, thereby reducing the workload of radiologists and mitigating the risk of human error.

Keywords: Radiology report generation · Multimodal learning · Knowledge graphs

1 Introduction

Radiology is a vital component of modern medical practice and plays a crucial role in medical diagnosis, treatment planning, and monitoring of diseases, which involves the interpretation of various medical imaging modalities such as X-rays, CT scans and MRI. From detecting fractures and tumors to monitoring the progression of chronic diseases, radiology is integral to both acute and long-term patient care. The current radiology report generation process predominantly relies on radiologists' verbal descriptions and subjective interpretations based on radiographs, which remains a largely manual and labor-intensive process. Radiologists often face a substantial workload and the generation of detailed accurate

reports can be time-consuming. This situation is amplified by the global shortage of skilled radiologists and the variability in report quality due to human factors such as fatigue and cognitive biases [22]. Errors or inconsistencies in radiology reports can have significant consequences on patient care and outcomes, which drive the need for accurate and automated solutions for radiology report generation (RRG), a complex cross-modal text generation task.

The integration of artificial intelligence (AI) in medical imaging has revolutionized the field of radiology. In recent decades, medical image analysis has achieved tremendous advancements in a variety of tasks such as classification and segmentation enhancing the ability of clinicians to interpret complex imaging data with greater accuracy and efficiency, which is primarily driven by the rapid development of deep learning [3]. Recently, vision-language pre-training approaches such as CLIP [21] and BLIP [11, 10] have achieved tremendous success on various multimodal downstream tasks including image caption [23], visual question [17], etc. Similarly, automated RRG aims to generate descriptive text from a set of radiographs. The current state-of-the-art deep learning models for RRG are built on the multimodal (vision-to-language) encoder-decoder architecture. However, they are still subject to several significant challenges when applied to RRG. One of the key challenges is that radiology reports are often rich in specialized complex medical terms, abbreviations and expressions that describe the anatomy and any abnormalities or changes observed, and sometimes suggest potential diagnoses. This requires that the deep learning models are extensively trained on medical texts. Domain-specific knowledge is crucial for accurate and contextually relevant radiology report generation. This gap highlights the need for a specialized approach that combines the strengths of advanced vision-language multimodal models and domain-specific knowledge in radiology. Knowledge Graphs (KGs) are designed to accumulate and convey factual knowledge of the real world, which organize and represent knowledge in the form of graphs, where the nodes represent entities of interest or attributes and edges represent relations between these entities and attributes [20]. In the medical domain, KGs have emerged as a powerful tool in organizing and representing complex healthcare data and medical knowledge. RadGraph is a knowledge graph dataset, whose entities and relations are extracted from full-text radiology reports [7]. One notable feature of radiology reports is their highly structured nature. Typically, physicians adhere to specific sentence patterns when describing diseases and organs, and employ consistent templates to construct the entire report [15]. Leveraging relevant entities and relations, LLM-augmented KG-to-text methods [19] can generate high-quality texts that accurately and consistently describe the knowledge graph information.

In this research, we explore multimodal learning approaches for learning joint embedding of modalities including images, graphs and texts. Leveraging existing medical and radiological knowledge graphs, we propose a knowledge graph-enhanced vision-to-language multimodal model for radiology report generation. This research aims to contribute meaningfully to Semantic Web research, showcasing its potential to enhance the semantic richness of domain-specific knowl-

edge for deep learning model training. Central to this endeavor is the utilization of medical knowledge graphs that link from various sources such as clinical data, research, drug information, etc., and provide a comprehensive understanding of domain knowledge.

2 State of the Art

Contrastive pre-training has received tremendous success in multimodal learning, especially in vision-language pretraining. Contrastive Language–Image Pre-training (CLIP) [21] involves a simple pre-training task that leverages the vast amount of text paired with images available on the internet. Utilizing loss such as InfoNCE [18], CLIP pulls image embedding and corresponding text embedding closer and pushes unpaired image and text farther in the embedding space. It can scale to achieve competitive zero-shot performance across a variety of image classification datasets. Following the idea of CLIP [21], You et al. [28] propose CXR-CLIP multi-view supervision [13] combining with image contrastive loss and text contrastive loss for enhancing the learning of study-level features of medical images and reports. Similarly, Li et al. [12] adopt BLIP [11, 10] for image-language multimodal learning, which employs a multimodal mixture of encoder-decoder architecture and utilizes the image-text contrastive loss to align the vision and language representations, image-text matching loss to learn cross-attention features between positive and negative image-text pairs, and finally language modeling loss for generating reports.

Knowledge injection approaches have recently emerged to address the lack of domain-specific knowledge and hallucination problems of language models [19]. Xu et al. propose KILM that injects entity-related knowledge into encoder-decoder pre-trained language models by incorporating knowledge-infilling into training objectives [26]. Zhang et al. propose GreaseLM layers that fuse encoded representations from pre-trained language models and graph neural networks over modality interaction operations [29]. Liu et al. utilize a pre-constructed knowledge graph as posterior knowledge to enhance RRG [16]. To address the limitation of approaches based on predicted knowledge graphs such as incomplete information, Li et al. [12] utilize a pre-constructed knowledge graph to assist the report generation process, however, their approach faces knowledge noises involved during the dynamic graph construction process caused by retrieved reports.

3 Problem Statement and Contributions

3.1 Problem Statement

The overarching goal of this research is to develop a model capable of generating accurate, coherent, and clinically relevant radiology reports from radiographs. The central problem to be addressed in this research is the lack of domain-specific knowledge of current state-of-the-art deep learning models. In this research,

we leverage existing knowledge graphs such as SNOMED CT and RadGraph and explore the concept of knowledge graph-enhanced vision-to-language multimodal models specifically designed for radiology report generation. Incorporating advanced AI techniques, particularly those involving multimodal learning and knowledge fusion, promises to not only streamline this process but also enhance the consistency and quality of radiology reports.

Research Questions To address this problem, the research is guided by the following questions:

- RQ1: How should we train the multimodal models that learn a joint embedding across different modalities, i.e., images, graphs and text?
- RQ2: How can we extract knowledge from the visual representation of radiographs?
- RQ3: How can we model and fuse the knowledge for enhanced report generation?

Research Hypothesis Based on these questions, the research hypothesis is formulated as follows: "*The integration of knowledge graph into the transformer-based vision-to-language multimodal model can improve the accuracy, consistency and clinical relevance of automated radiology report generation*". This hypothesis is based on the assumption that transformer-based crossmodal models are capable of synergizing the embeddings of visual features (radiology images) and structured knowledge (ontologies, medical lexicons, etc.) to produce more accurate, consistent, semantic richness, and clinically relevant radiology.

3.2 Contributions

A novel aspect of this research is the integration of semantic web technologies, in particular knowledge graphs, into the multimodal deep learning model, which enhances the model's capacity to utilize complex medical terminologies and relations and enrich the semantic quality of radiology reports. A significant part of this research will involve a thorough empirical evaluation of the model against traditional and existing automated methods, focusing on clinical relevance and accuracy. By applying knowledge graphs to a practical healthcare domain, this research aims to contribute meaningfully to Semantic Web research, showcasing its potential to enhance data interpretation and utility in medical contexts.

4 Research Methodology and Approach

After reviewing relevant literature, we have identified the research gap we are going to tackle in this research, i.e., the lack of domain-specific knowledge in the current deep learning-based approaches, which also leads to the formulation of research questions to guide the research. In this section, we outline the

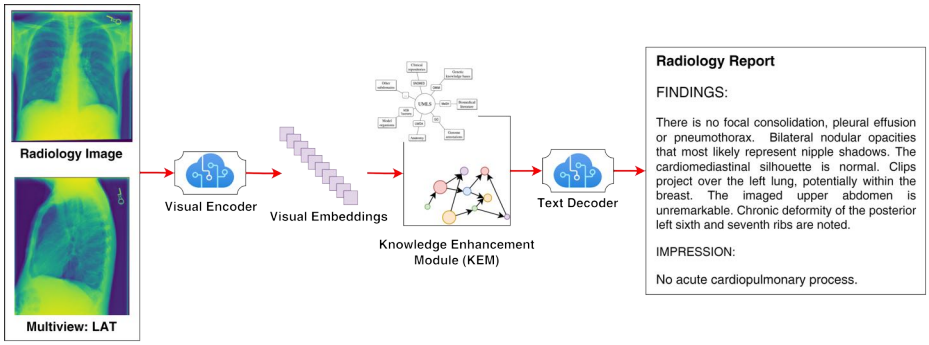

Fig. 1. Architecture overview of knowledge-enhanced image-to-language multimodal model for radiology report generation.

methodology employed in developing the proposed knowledge graph-enhanced image-to-language multimodal model for radiology report generation. Drawing from insights garnered during the literature review, we design our architectural framework depicted in Figure 1. Central to this architecture is the incorporation of a Knowledge Enhancement Module (KEM) as shown in Figure 2, aimed at extracting and integrating relevant domain knowledge into the model.

4.1 Architecture Overview

As shown in Figure 1, we adopt an encoder-decoder architecture that is widely used in current state-of-the-art approaches [15] and introduce a Knowledge Enhancement Module (KEM) for knowledge extraction, processing and fusion. Formally, the model employs a visual encoder f_v to extract visual features $\mathcal{H}^v = f_v(\mathcal{I})$ from a radiograph $\mathcal{I}$. The KEM module takes visual embedding as input and extracts, processes and fuses the knowledge into the joint embedding $\mathcal{K} = KEM(H^v)$. Finally, a text decoder f_t translates the fused features into report texts $\hat{\mathcal{R}} = f_t(\mathcal{K})$. Deep learning-based approaches utilize radiology reports $\mathcal{R}^*$ written by professional radiologists as the reference for the corresponding radiographs and the objective is to minimize the discrepancies of the output of the model $\hat{\mathcal{R}}$ and the reference report $\mathcal{R}^*$.

4.2 Multimodal Model Training

To address RQ1, we will investigate various training strategies to optimize the effectiveness of multimodal models for radiology report generation. We leverage recent advancements in multimodal learning and contrastive learning techniques, such as [30, 5], to learn a joint embedding across different modalities, i.e., images, graphs and text, and to bridge the modality gap, enhancing the interchangeability of embeddings. Contrastive learning can drive a variety of pretext tasks, however, most studies follow instance discrimination tasks, which consider

a query and a key as a positive pair if they originate from the same image-text pair, and otherwise as a negative sample pair. However, image-text pairs in the radiology report dataset could be correlated as the same disease causes the same symptom and observation. More effective contrastive learning approaches for multimodalities, including image, graph and text, need to be explored.

4.3 Knowledge Enhancement Module

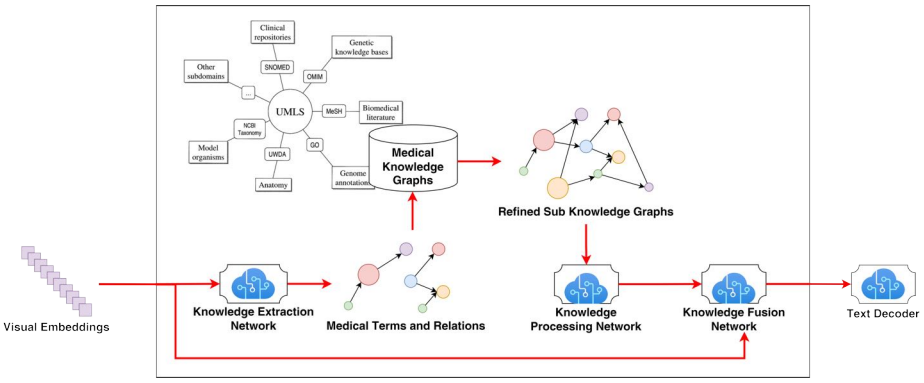

Fig. 2. Knowledge Enhancement Module (KEM) for knowledge extraction, processing and fusion.

To enrich the semantics of the extracted visual features, we propose the knowledge enhancement module for extraction, processing, and fusion of domain-specific knowledge as shown in Figure 2.

Sub-knowledge Graph Construction and Refinement A knowledge extraction network f_{ke} is employed to predict included knowledge $\mathcal{G}^m = f_k(H^v)$ such as medical terms and relations, then the extracted knowledge is refined and used to construct a sub knowledge graph $\mathcal{G}_{sub}$ from large scale medical knowledge graphs such as UMLS. To address RQ2, we will employ various techniques for extracting domain-specific knowledge from the visual representation of radiographs. This may include but is not limited to the utilization of attention mechanisms to identify salient regions and features within the radiographs and the incorporation of domain-specific rules and heuristics to guide the feature extraction process.

Knowledge Processing and Fusion Addressing RQ3 involves devising effective strategies to model and fuse the extracted knowledge (in the form of graphs) for enhanced report generation. Traditional embedding methods can not well model such kind of multi-relational data. We employ the variants of graph

attention network [24] from [27] for modeling and processing knowledge graphs. Similar to GreaseLM [29], we utilize a cross-modal fuser $\mathcal{K} = f_f(\mathcal{H}^v, f_g(\mathcal{G}_{sub}))$ to integrate the embeddings of visual and knowledge graph features.

5 Evaluation Plan

For the validation of the hypothesis and the evaluation of the effectiveness of our proposed approaches for RRG, we describe our evaluation plan in this section, which is designed to measure the accuracy, consistency, and clinical relevance of the reports generated by our proposed models, comparing them with other state-of-the-art models.

5.1 Datasets

The success of any deep learning model heavily relies on the quality and quantity of the data used for training and testing. In this research, we will collect a diverse dataset of radiographs along with their corresponding expert-generated radiology reports or annotations, as well as medical knowledge graphs that models acquire domain-specific knowledge.

Radiograph Datasets Essential to our model's training and evaluation are extensive datasets comprising radiographs and corresponding reports. The datasets selected include MIMIC-CXR [8, 9], IU X-Ray [4], NIH ChestX-ray [25], Chexpert [6], and for multilingual capabilities, datasets like CX-CHR [14] and PadChest [2].

Medical Knowledge Graphs Our models not only rely on image and text data but also integrate substantial medical knowledge from various sources, such as knowledge graphs. The integration of knowledge graphs aims to enhance the clinical accuracy and relevance of the generated reports. The key knowledge graphs we are incorporating are UMLS [1] and RadGraph [7].

5.2 Metrics

Our evaluation plan employs a multifaceted set of metrics to measure the accuracy, consistency, and clinical relevance of the reports generated by our models, offering a comprehensive comparison with traditional manual methods and state-of-the-art models. In general, our chosen metrics for RRG are categorized into three types based on the target granularity, namely, entity level, graph level, and report level. The evaluation metrics are detailed as follows:

1. **Entity Recognition and Graph Construction**: Metrics such as accuracy, recall, precision, and F_1 score are used to evaluate the clinical entity recognition. To ensure the clinical utility of the generated reports, we also

calculate the accuracy of the diagnoses provided in the generated reports compared to the diagnoses concluded from the reference standard. We evaluate graph construction by using metrics including F_1 score and ROUGE-2, similar to the Tianchi knowledge graph construction competition[1].

2. **Report Accuracy and Consistency**: We use metrics from natural language generation and image caption tasks to evaluate the report accuracy and consistency compared to the reference reports, including BLUE, METEOR, ROUGE, CIDEr and BERTScore.

Besides, we plan to conduct a proper user study in collaboration with the radiology department and pathology department at University Hospital Aachen, as those metrics from traditional natural language processing tasks still have their limits and do not provide qualitative results. Expert assessment of the generated reports can be more reasonable but expensive and time-consuming.

6 Results

As this paper is part of an early-stage Ph.D. symposium, we are currently in the process of running baseline models. Consequently, we do not yet have any results to report at the time of writing. However, we can refer to the results in reference [15], which indicate that knowledge enhancement and cross-modal approaches have a positive impact compared to image caption methods.

7 Conclusions/Lessons Learned

In conclusion, this paper has explored the challenges and limitations that current approaches are facing for radiology report generation, highlighting the critical issue of the lack of domain-specific knowledge in existing vision-language multimodal models. To address this research gap, we propose the integration of a knowledge enhancement module into existing vision-language multimodal models. The KEM module aims to extract and incorporate domain-specific knowledge from the visual embedding of radiograph features and medical knowledge graphs, enhancing the model's understanding of medical concepts and improving the quality of generated reports. By leveraging advanced AI techniques such as multimodal learning and knowledge fusion, the proposed approach seeks to overcome the limitations of current methods and pave the way for more effective and clinically useful radiology report generation systems. However, since this research is at an early stage, we are looking forward to discussing the research challenges we try to address in this paper and other challenges we may not be aware of, possible improvements to our proposed architecture and modules (Figures 1 and 2), as well as the implementation of an effective plan to evaluate our research.

[1] https://tianchi.aliyun.com/competition/entrance/532080/information

References

1. Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), D267–D270 (2004)
2. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
3. Chen, X., Wang, X., Zhang, K., Fung, K.M., Thai, T.C., Moore, K., Mannel, R.S., Liu, H., Zheng, B., Qiu, Y.: Recent advances and clinical applications of deep learning in medical image analysis. Medical Image Analysis **79**, 102444 (2022)
4. 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)
5. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (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. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)
9. 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)
10. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
11. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
12. Li, M., Lin, B., Chen, Z., Lin, H., Liang, X., Chang, X.: Dynamic graph enhanced contrastive learning for chest x-ray report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3334–3343 (2023)
13. Li, Y., Liang, F., Zhao, L., Cui, Y., Ouyang, W., Shao, J., Yu, F., Yan, J.: Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. arXiv preprint arXiv:2110.05208 (2021)
14. 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)

15. Liu, C., Tian, Y., Song, Y.: A systematic review of deep learning-based research on radiology report generation. arXiv preprint arXiv:2311.14199 (2023)
16. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2021)
17. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
18. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
19. Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., Wu, X.: Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering (2024)
20. Peng, C., Xia, F., Naseriparsa, M., Osborne, F.: Knowledge graphs: Opportunities and challenges. Artificial Intelligence Review pp. 1–32 (2023)
21. 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)
22. Singh, S., Karimi, S., Ho-Shon, K., Hamey, L.: Show, tell and summarise: learning to generate and summarise radiology findings from medical images. Neural Computing and Applications **33**, 7441–7465 (2021)
23. Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., Cucchiara, R.: From show to tell: A survey on deep learning-based image captioning. IEEE transactions on pattern analysis and machine intelligence **45**(1), 539–559 (2022)
24. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y., et al.: Graph attention networks. stat **1050**(20), 10–48550 (2017)
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
26. Xu, Y., Namazifar, M., Hazarika, D., Padmakumar, A., Liu, Y., Hakkani-Tür, D.: Kilm: Knowledge injection into encoder-decoder language models. arXiv preprint arXiv:2302.09170 (2023)
27. Yasunaga, M., Ren, H., Bosselut, A., Liang, P., Leskovec, J.: Qa-gnn: Reasoning with language models and knowledge graphs for question answering. arXiv preprint arXiv:2104.06378 (2021)
28. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
29. Zhang, X., Bosselut, A., Yasunaga, M., Ren, H., Liang, P., Manning, C.D., Leskovec, J.: Greaselm: Graph reasoning enhanced language models. In: International conference on learning representations (2021)
30. Zhang, Y., Sui, E., Yeung-Levy, S.: Connect, collapse, corrupt: Learning cross-modal tasks with uni-modal data. arXiv preprint arXiv:2401.08567 (2024)