

Medical Report Generation: A Hierarchical Task Structure-Based Cross-Modal Causal Intervention Framework

Yucheng Song

School of Computer Science and Engineering
Central South University
Changsha 410083, China

Yifan Ge

Institution2
School of Computer Science and Technology
University of Science and Technology of China
Anhui 230027, China

Junhao Li

School of Computer Science and Engineering
Central South University
Changsha 410083, China

Zhining Liao

Glasgow Lab for Data Science & AI
Public Health, School of Health & Wellbeing
University of Glasgow
Glasgow, UK

Zhifang Liao

School of Computer Science and Engineering
Central South University
Changsha 410083, China

Abstract

Medical Report Generation (MRG) is an indispensable component of modern medical diagnostics. It can automatically generate corresponding medical reports based on given radiological images, alleviating the burden on radiologists. However, designing a MRG model capable of reliably describing lesion areas still presents the following challenges: 1) insufficient understanding of domain-specific knowledge; 2) poor embedding space alignment between entities in the report text and visual signals; and 3) spurious correlations arising from visual and linguistic biases. Previous attempts have mostly focused on addressing individual challenges without comprehensively considering the issues faced by MRG. In this paper, we aim to tackle the above three formidable challenges through a novel hierarchical task decomposition perspective. To this end, we propose a Hierarchical Task Structure-Based Cross-Modal Causal Intervention Framework (HTSC-CIF) for MRG. This framework is the first to classify the three challenges of MRG into different levels. In the low-level task, we adopt a novel method for aligning entity features specific to medical tasks with spatial locations to help the upper-level tasks' visual encoder better understand domain knowledge in terms of

spatial positions. In the mid-level task, we employ Prefix Language Modeling for text and Masked Image Modeling for images, enhancing cross-modal understanding and alignment through mutually guided generation. Finally, to mitigate the spurious correlations caused by cross-modal data biases, we design a cross-modal causal intervention module. This module aims to reduce cross-model confounders through causal front-door intervention, thereby achieving enhanced interpretability in the high-level task. Extensive experimental results demonstrate that our hierarchical task framework is reasonable and effective, significantly outperforming other state-of-the-art methods in the MRG task. We make code public on the acceptance of the paper.

1. Introduction

In the realm of medical imaging technology, the identification of latent pathological changes and the formulation of lucid diagnostic reports represent a time-consuming and technically demanding task. This imposes a significant burden on radiologists, particularly compromising the quality of reports and elevating the potential for errors [9]. This

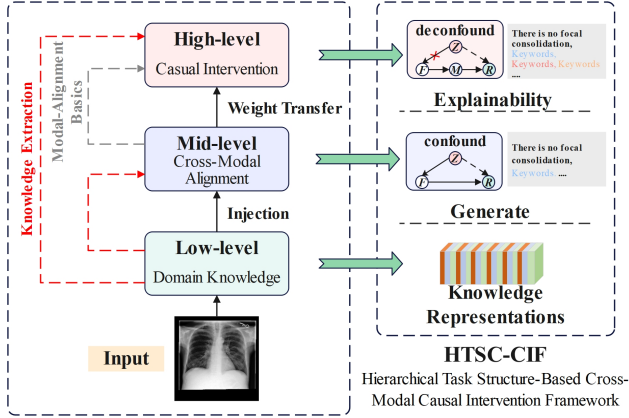


Figure 1. Multi-level task design of HTSC-CIF.

challenge necessitates the exploration of automated Medical Report Generation (MRG) systems. However, current MRG also faces several challenges: 1) How to incorporate rich domain knowledge into the model to improve the accuracy and reliability of the report. Medical images contain a significant amount of specialized information, which often requires a deep medical background to fully comprehend; 2) How to effectively optimize visual encoding and ensure the spatial alignment between entity descriptions in the report text and visual signals. In medical image analysis, visual encoding must not only capture the visual features in the image but also ensure that these features accurately correspond to the text descriptions in the report; 3) Spurious correlations caused by visual and linguistic biases. This bias may arise from inconsistencies between the image data and the statistical descriptions in the text, leading to misleading associations during model training. Specifically, the model may incorrectly associate certain visual features with specific statistical characteristics of text descriptions, despite these associations lacking any medical basis. This phenomenon not only reduces the accuracy of the report but also significantly impacts the interpretability of the model.

To address the above challenges, researchers have made significant efforts to advance MRG methods [11]. Firstly, existing studies integrate domain knowledge to strengthen visual encoding and contextual understanding, thereby laying the foundation for cross-modal alignment [14, 19, 21, 39]. Simultaneously, researchers explore cross-modal alignment between visual signals and entity descriptions to achieve semantic consistency [1, 22, 25, 35]. However, due to spurious correlations in image-text data arising from visual and linguistic biases, generating accurate and reliable reports that describe lesion regions remains highly challenging. To resolve this, many studies incorporate causal intervention methods into MRG to enhance report accuracy and interpretability [2, 3]. It is evident that these works are interconnected, progressively deepening, and mutually

reinforcing. Therefore, an effective approach is needed to combine domain knowledge integration, cross-modal alignment of visual and textual features, and causal intervention, thereby comprehensively improving the accuracy, reliability, and interpretability of medical imaging diagnostic report generation.

In this paper, we propose a Hierarchical Task Structure-Based Cross-Modal Causal Intervention Framework (HTSC-CIF) for medical imaging report generation. This framework primarily consists of three hierarchical tasks: domain knowledge enhancement, cross-modal alignment, and interpretability enhancement under causal intervention. These tasks can be regarded as low-level, mid-level, and high-level tasks in MRG. For the low-level task, we design a domain knowledge enhancement module. By employing a novel entity feature description and spatial alignment method tailored to specific medical tasks, this module assists the visual encoder in better understanding domain knowledge at the spatial level for higher-level tasks. For the mid-level task, we design a cross-modal alignment module. We employ Prefix Language Modeling (PLM) for text and Masked Image Modeling (MIM) for images, enhancing cross-modal alignment through guided generation. This approach enables pretraining a network capable of generating preliminary report results. Finally, to mitigate spurious correlations caused by cross-modal data bias, we transfer the weights of the mid-level pre-trained network and design a cross-modal causal intervention module. This module aims to reduce cross-modal confounders through causal front-door intervention, thereby achieving the high-level task of interpretability enhancement. To the best of our knowledge, our work is the first to address the three challenges of medical image report generation. The main contributions of this paper are summarized as follows:

- We are the first to simultaneously perform low-level domain knowledge enhancement, mid-level cross-modal information alignment, and high-level causal intervention tasks in MRG. These three tasks mutually guide and progressively promote each other from low to high levels.
- We achieve the injection enhancement of low-level domain knowledge into upper-level tasks, and the upper-level tasks can obtain rich domain knowledge without additional the network architecture.
- We experimentally demonstrate the superiority of HTSC-CIF and SOTA performance on two public benchmark datasets.

2. Related Work

2.1. Domain Knowledge Enhancement

Enhancing domain knowledge in medical image report generation improves model understanding of image data, leading to more accurate, clinically relevant, and interpretable

reports. Researchers have developed various injection methods for MRG [41], including using knowledge graphs [44, 51] to reduce language biases and fuse visual features [42]. Other approaches extract medical concepts from reports [50] to enrich semantics. As a crucial foundational task, domain knowledge provides medical context, and we optimize the visual encoder by injecting it to better capture image features, enhancing the ability to parse data for report generation.

2.2. Cross-Modal Alignment

Cross-modal alignment in medical image report generation ensures semantic consistency between visual and textual data, improving report accuracy by addressing modality differences. Key methods include alignment modules for embeddings [28], global/local alignment with reconstruction [8], contrastive loss for global alignment [20], and cross-modal networks for feature embedding [7, 33]. As a mid-level task, it reduces modal bias and enhances clinical utility in MRG.

2.3. Causal Intervention

Existing MRG methods focus on knowledge embedding and modal alignment but often overlook visual-linguistic bias, which causal reasoning addresses by eliminating spurious correlations through interventions [18, 31, 40]. Causal inference mitigates confounding via front-door interventions [3], though limitations in handling unobservable factors motivate low-level domain knowledge injection. Positioned as a high-level task, it relies on foundational knowledge and mid-level alignment to accurately process medical contexts, reduce ambiguities, and enhance MRG’s robustness and interpretability.

3. Method

3.1. Framework

Research shows that knowledge enhancement, modality alignment, and interpretability enhancement in MRG are closely related to each other. Therefore, it is intuitive to learn these three tasks together under a unified framework. This paper proposes a new cross-modal causal intervention framework for MRG, called the Hierarchical Task Structure-Based Cross-Modal Causal Intervention Framework (as shown in Figure 2). The framework consists of three main parts: domain knowledge enhancement module, cross-modal alignment module, and cross-modal causal intervention module.

3.2. Domain Knowledge Enhancement Module

In medical report generation tasks, domain knowledge plays a critical role [23, 45]. This knowledge primarily derives from entity information in medical reports, including professional medical concepts such as disease categories (e.g.,

pneumonia, pleural effusion) and affected anatomical regions (e.g., left upper lung zone) [46]. Such entity information can serve as supervisory signals to optimize the learning process of visual encoders. To make use of these entities, we propose an Entity Contrastive Loss Optimization (ECLO) method based on entity extraction.

The design of ECLO aims to guide and optimize the model’s learning and alignment of image features through a contrastive learning strategy, leveraging entity information extracted from medical text reports. This module minimizes the difference between image features and text descriptions by applying a weighted combination of contrast loss based on entity location prediction and binary cross entropy loss, thereby improving the model’s recognition and positioning accuracy of medical entities. Where, the binary cross entropy loss L_{cls} is used for entity existence prediction. It measures the discrepancy between the probability of the model predicting the existence of an entity and the true label, which can be defined as:

$$L_{cls} = - \sum_{k=1}^N y_k \log \hat{s}_k \quad (1)$$

The contrastive loss based on entity location prediction is used for the prediction of entity location. It calculates the difference between the model’s predicted probability of entity existence and the true label, which can be defined as:

$$L_{loc} = - \frac{1}{Q} \sum_{K=1}^{|Q|} \left(\frac{e^{\langle \hat{p}_k, p_k \rangle}}{e^{\langle \hat{p}_k, p_k \rangle} + \sum_{u=1}^M e^{\langle \hat{p}_k, P_{I(k,u)} \rangle}} \right) \quad (2)$$

Where, $\langle \cdot, \cdot \rangle$ represents the dot product of the vectors, $|Q|$ is the total number of queries, p_k is the entity existence prediction, $\hat{p}_k$ is the model predicted position embedding, p_k is the true position embedding, M is the number of negative samples, and $I(k, u)$ is a random index sampling function used to sample negative samples from the position set P . Through the domain knowledge enhancement module, the model can simultaneously learn to predict the existence of entities and their precise locations within the images, thereby optimizing the embedding capabilities of the visual encoder for mid-level and high-level tasks. In this stage, the optimization objective can be expressed as minimizing the entity existence prediction loss L_{cls} and the entity location prediction loss L_{loc} . Therefore, the loss can be defined as:

$$L_{low} = L_{cls} + L_{loc} \quad (3)$$

3.3. Cross-Modal Alignment Module

In the low-level tasks, the embedding capabilities of the visual encoder are optimized through domain knowledge enhancement. However, during the process of cross-modal

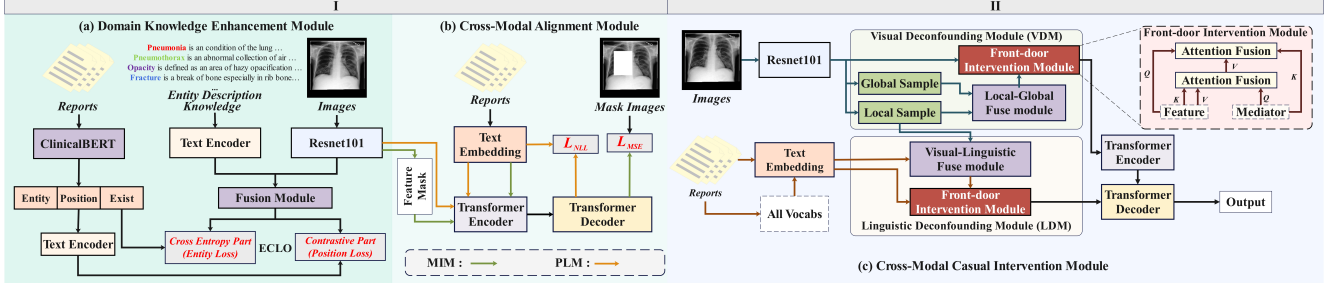


Figure 2. The overall structure of HTSC-CIF. (a) Domain Knowledge Enhancement Module. (b) Cross-Modal Alignment Module. (c) Cross-Modal Causal Intervention Module.

information alignment and understanding in the mid-level task, significant semantic discrepancies still persist. Differences in expression and comprehension between different modalities (e.g., image data and text data) may lead to modal biases, even after the visual encoder has been optimized with domain knowledge. This can result in inconsistencies between visual features and textual descriptions. Inspired by SimVLM [36] and VICI [3], we used Prefix Language Modeling (PLM) and Mask Image Modeling (MIM) to address this issue.

3.3.1. Prefix Language Modeling.

In PLM, the report text is randomly divided into two parts, and the key idea is to use one part to guide the generation of the other part. We feed the visual features $F_v \in R^{\frac{H \times W}{P^2} \times d}$ extracted from the original image into the Transformer encoder, where P is the patch size and d is the embedding dimension, and use them along with the prefix text description $F_{w < n_p}$ to guide the generation of the suffix text $w_{n_p}, \dots, w_n$. To enhance the generalizability of the method, in the absence of visual features, the prefix text can also be used directly to generate the suffix text. The loss function for this part can be defined as:

$$L_{PLM} = - \sum_{i=n_p}^n \log P_{\theta}(w_i | F_v, F_{w < n_p}) \quad (4)$$

Where, θ represents the trainable parameters of the model, F_v represents the visual features with trainable 2D position encoding, $F_{w < n_p}$ is the prefix text embedding, and n represents the length of the report.

3.3.2. Masked Image Modeling.

The MIM facilitates better alignment between modalities by reconstructing masked visual features. Inspired by MAE [13] and considering the specific requirements of the image report generation task, we use both image and text modalities to reconstruct the masked visual features. Specifically, we use a CNN to extract features from the original image to obtain low-resolution features, and the original high-resolution image is reconstructed through a combination of

the unmasked low-resolution visual features and text embeddings. The loss function for this part can be defined as:

$$L_{MIM} = P_{\theta}(F_{vm} | F_{vum}, F_w) \quad (5)$$

Among them, the visual features are extracted from the Resnet101 backbone after the optimization of the low-level task, F_{vm} represents the masked visual features, F_{vum} represents the unmasked low-resolution visual features, and F_w represents the text embedding of the complete report. The above two parts further enhance the consistency between the visual features and the text description, thereby realizing the pre-training modeling of report generation under cross-modality guidance.

Unlike low-level tasks, mid-level and high-level tasks can independently complete the report generation task. However, their optimization objectives differ. For mid-level tasks, the optimization objective can be expressed as minimizing the Masked Image Modeling (MIM) loss and the Prefix Language Modeling (PLM) loss. Therefore, the loss can be defined as:

$$L_{mid} = L_{MIM} + L_{PLM} \quad (6)$$

3.4. Cross-Modal Causal Intervention Module

After pre-training in cross-modal understanding through the mid-level task, the model has acquired a certain level of joint modal comprehension capability. However, the fusion process of heterogeneous modalities may still be influenced by cross-modal confounding factors [47]. To address this issue, we employ the front-door causal intervention method to mitigate the visual-language bias arising from the heterogeneity in multimodal data, aiming to enhance the interpretability of reports generated by the Hierarchical Task Structure-Based Cross-Modal Causal Intervention Framework.

3.4.1. Causal Structural Modeling.

We employ the Structural Causal Model (SCM) to represent the entire inferential process. In the field of causal inference, we can achieve the elimination of confounding

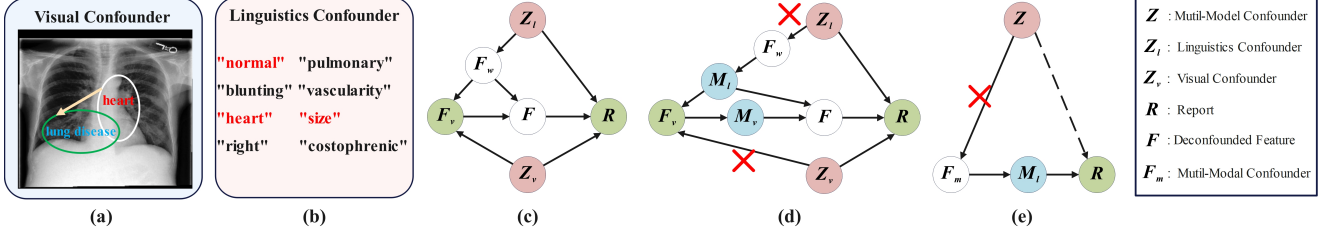


Figure 3. Description of causal structural modeling.

effects through front-door intervention and back-door intervention. Specifically, the chain structure $X \rightarrow Y$ indicates a causal effect of X on Y , expressed as $P(Y|X)$. If the confounding factor Z is influencing the relationship, this can be represented as $X \leftarrow Z \rightarrow Y$ (visually, as shown in Figure 3(a), the heart and the lungs have overlapping areas in the region. These overlapping areas serve as visual confounding factors and jointly affect the analysis results of the disease in a certain area of the image. In text, as shown in Figure 3(b), the description of the same medical image, different organs may share some terminology, for example, the description of organ A will affect the reasoning analysis of organ B). The *do* calculus can be introduced and the back-door causal intervention can be used to condition on the X variable. The result Y can be defined as:

$$P(Y|do(X)) = \sum_Z P(Y|X, Z = z)P(Z = z) \quad (7)$$

Compared with back-door intervention, front-door intervention provides an implicit way to deconfound. To eliminate unobservable confounding effects, an mediator M is introduced to cut off the chain structure $X \leftarrow Z \rightarrow Y$. Here, the mediator M is introduced by X and there is no back-door path. There is no direct causal path between X and Y , so our intervention on X can be defined as:

$$P(Y|do(X)) = \sum_m P(Y|do(M = m))P(M = m|X = x) \quad (8)$$

Now, to estimate $P(Y|do(M = m))$, we can use back-door intervention to cut off the chain structure $M \leftarrow X \leftarrow Z \rightarrow Y$, which can be defined as:

$$P(Y|do(M = m)) = \sum_{\hat{x}} P(X = \hat{x})P(Y|X = \hat{x}, M = m) \quad (9)$$

Where, $\hat{x}$ comes from the input X , not from the mediator M . Finally, we can apply Equation 8 to Equation 9 and obtain:

$$\begin{aligned} P(Y|do(X)) &= \sum_m P(M = m|X = x) \sum_{\hat{x}} P(X = \hat{x})P(Y|X = \hat{x}, M = m) \end{aligned} \quad (10)$$

We construct a structural causal model from the causal paths of two different modalities, vision and language, as shown in Figure 3(c). Based on Equation 10 derived above, the causal relationships $F_v \rightarrow R$ and $F_w \rightarrow R$ are affected by the confounding factors $Z = Z_v, Z_l$ from the back-door paths $F_v \leftarrow Z_v \rightarrow R$ and $F_w \leftarrow Z_l \rightarrow R$ respectively. In SCM, non-interventional prediction can be defined as:

$$\begin{aligned} P(R|I) &= P(R|F_v, F_w) \\ &= \sum_{i=1}^n \sum_z P(w_i|F_v, F_w, Z = z)P(Z = z|F_v, F_w) \end{aligned} \quad (11)$$

Where, Z will introduce many spurious correlations into the generated report, and it is necessary to introduce a mediator and perform front-door intervention, as shown in Figure 3(d). Z_v as a visual confounding factor is difficult to observe, but the back-door path $F_v \leftarrow Z_v \rightarrow R$ can be cut off by learning the true causal relationship $F_v \rightarrow M_v \rightarrow F \rightarrow R$. Similarly, in the back-door path $F_w \leftarrow F_w \rightarrow Z_l \rightarrow R$, the back-door path can also be cut off by calculating M , which can be defined as:

$$P(R|do(I)) = P(R|do(F_v), do(F_w)) \quad (12)$$

$$P(R|do(F_v), do(F_w)) \approx \text{Softmax}(g(F_w, F_v, \hat{M}_v, \hat{M}_l)) \quad (13)$$

Where, Equation 11 uses the Normalized Weighted Geometric Mean (NWGM) method [43] for calculation. $g(\cdot)$ represents the complete deep learning mapping process of the Vision-Language Model. Figure 3 (e) shows the overall structure after causal intervention. After we model the causal structure, we need to use the deconfounding module to eliminate the confounding effect between modalities.

3.4.2. Visual Deconfounding Module.

In the Visual Deconfounding Module, we implement the visual mediator M_v , which is composed of a local feature F_{vl} and a global feature F_{vg} . Inspired by TransFG [12], we select the first k tokens with high attention accumulation as the core visual representation of the corresponding area in the report, denoted as $F_{vl} = R^{k \times d}$, where d is the dimension of the Transformer. On this basis, we use Caam [34] to further enhance F_{vl} . F_{vg} is implemented through global sampling. We use a down-sampled transformer module to convert the visually encoded 14×14 tokens into 7×7 tokens, which is $F_{vg} = R^{49 \times d}$, and adopts maximum pooling to better ensure the global structural information, the equation can be defined as:

$$F_{vg} = L[MP(F_v) + Attn(MP(LN(F_v)))] \quad (14)$$

Where, MP represents the two-dimensional maximum pooling, $Attn$ is the two-dimensional correlation attention, LN is the layer normalization, and L is the linear layer. After completing the calculation of F_{vl} and F_{vg} , we use multi-head attention to fuse the two parts of the features to obtain the mediator M_v , the equation can be defined as:

$$M_v = FFN([MHA(F_{vl}, F_{vl}, F_{vl}), MHA(F_{vl}, F_{vg}, F_{vg})]) \quad (15)$$

Where, MHA is the multi-head attention layer, FFN is the feed-forward neural network layer, and $[\cdot, \cdot]$ here represents the concatenation.

3.4.3. Language Deconfounding Module.

In the Language Deconfounding Module, we implement the linguistic mediator M_l , which can be leveraged to block the backdoor path $F_v \leftarrow F_w \leftarrow Z_l \rightarrow R$. Reconstruct F_{vl} using the tokens of the words in the vocabulary, we get F'_{vl} , and then calculate M_l , we can introduce visual factors to reduce the dependence on word frequency in the generated word data, the equation can be defined as:

$$F'_{vl} = FFN(MHA(F_{vl}, \hat{W}, \hat{W})) \quad (16)$$

$$M_l = FFN(MHA(F'_{vl}, F_{vl}, F_{vl})) \quad (17)$$

Where, $\hat{w}$ represents the tokens of all words in the vocabulary. We establish a causal relationship $F_w \leftarrow M_l \rightarrow F_v \rightarrow M_v \rightarrow F \rightarrow R$ to cut off the back-door path $F_v \leftarrow F_w \leftarrow Z_l \rightarrow R$. The reconstructed visual mediator is input into the transformer decoder together with the language mediator to learn the fused cross-modal features.

The high-level task possesses independent causal intervention modules (LDM and VDM). The remaining modules share the pre-trained model weights with those in the

mid-level task and are fine-tuned on the pre-trained model. Since the high-level task only focuses on text generation, the optimization objective can be expressed as minimizing the negative log-likelihood loss L_{nll} of the generated text, which can be defined as:

$$L_{high} = L_{nll} = - \sum_{i=1}^n \log \left(\text{Softmax}(g(F_{w<i}, F_v, \hat{M}_v, \hat{M}_l)) \right) \quad (18)$$

Where, n is the length of the generated image report and $F_{w<i}$ is the prefix text when estimating the word w_i .

3.5. Training loss

Our training process can be divided into two stages. The first stage is the joint training of low-level and mid-level tasks. This stage aims to provide pre-trained weights for the shared modules in the high-level task, and the loss is defined as:

$$L_{stage-1} = \lambda L_{low} + (1 - \lambda) L_{mid} \quad (19)$$

Where λ is a hyperparameter in the first stage. Through experiments, we found that when $\lambda = 0.25$, the model achieved the best performance. The ablation studies on parameter selection can be found in the supplementary materials.

The second stage is the training of the high-level task. The purpose is to fine-tune based on the pre-trained weights and assign appropriate parameters to LDM and VDM on the basis of prior knowledge. The loss is defined as:

$$L_{stage-2} = L_{high} \quad (20)$$

4. Experiments & Results

4.1. Datasets.

We conduct our experiments on two conventional benchmark datasets, i.e., MIMIC-CXR [17] is the largest dataset in the field of medical image reports, containing 377,110 chest X-ray images and 227,835 paired texts. We use the official partitioning to obtain a training set of 368,960 images and 222,758 reports, a validation set of 2,991 images and 1,808 reports, and a test set of 5,159 images and 3,269 reports. IU-Xray [10] is the widely-used public benchmark dataset for medical report generation and contains 7,470 chest X-ray images associated with 3,955 fully de-identified medical reports. Each report is composed of impression, findings and indication sections, etc.

To provide fine-grained objects for domain knowledge enhancement, we utilized RadGraph developed by Jain et al. [15] to extract entity knowledge from the reports of the MIMIC-CXR dataset. However, since RadGraph has

not processed the IU-Xray dataset, we employed dygie++ [32] and AGXnet [49] to process IU-Xray, resulting in a new entity-relation dataset for IU-Xray. The specific methods are detailed in the supplementary materials, and we will make this dataset publicly available.

4.2. Evaluation Metrics.

We evaluate model performance using Natural Language Generation (NLG) metrics. NLG metrics include BLEU, METEOR, and ROUGE-L. On the other hand, CIDEr is more suitable for captioning tasks, as it emphasizes the importance of topic-specific terms (critical in the MRG task) while downweighting common phrases.

In the first stage, we use the first three layers of ResNet101 as the visual backbone to extract X-ray images with an input size of 224×224 . The output feature map is $V \in R^{14 \times 14 \times 512}$. In the low-level task, we use the pre-trained ClinicalBERT to complete entity and position embeddings, obtaining $|Q| = 75$ entities and the $|P| = 51$ most frequently occurring positions. For each entity, $M = 7$ negative positions are set for contrastive loss calculation. In the mid-level task, it shares the same visual encoding module as the low-level task. The embedding dimension of the transformer is 512, the number of heads is 8, the image mask rate of MIM is 85%. We use the AdamW optimizer, with the maximum learning rate set to $5e-4$, the weight decay set to $1e-2$, and the number of epochs set to 30.

In the second stage, we transfer the pre-trained weights obtained in the first stage and perform fine-tuning based on them. The shared modules maintain the same parameters as those in the mid-level task. The main objective is to train the causal intervention modules. We use the Adam optimizer, with the initial learning rate set to $1e-5$, the weight decay set to $5e-5$, and the number of epochs set to 10. All of our experiments are conducted on 4 GeForce RTX 4090 GPUs.

4.3. Results and Analyses

4.3.1. Comparative Experiment.

The comparative experimental results are shown in Table 1. Overall, on the MIMIC-CXR and IU-Xray datasets, our model achieved SOTA performance in BLEU-1, BLEU-3, BLEU-4, METEOR, and ROUGE-L, and the second-best performance in BLEU-2 and CIDEr. At the same time, our method also outperforms existing causal inference methods [4]. This demonstrates the effectiveness of the hierarchical approach, as the lower-level tasks have led to noticeable improvements in metrics for the middle-level and high-level tasks. In particular, the leading performance in METEOR indicates that the generated reports are more in line with the professional expression habits of clinicians. Although CIDEr is slightly inferior to the best model, the SOTA performance in ROUGE-L suggests that the model’s coverage

of the vast majority of key pathological terms is already highly comprehensive. The minor gap in CIDEr may only reflect the need for further optimization in generating a very small number of rare terms.

4.3.2. Ablation Experiment.

We further investigated the effectiveness of HTSC-CIF through ablation studies. In this section, we conducted two types of ablation experiments on the MIMIC-CXR dataset. The first type focused on exploring different combinations of tasks across various hierarchical levels. The second type examined the combinations of deconfounding modules within the high-level tasks.

We used a Transformer network as the Baseline and expanded the hierarchical levels based on it. The results are shown in Table 2. The results indicate that tasks at each hierarchical level are helpful for MRG, and the final HTSC-CIF achieved the best performance in terms of metrics. Additionally, an interesting phenomenon was observed: the ROUGE-L metric showed the best performance when only high-level tasks were added. This is because a complex model may be highly consistent with the reference text semantically, but the generated text may use different expressions, resulting in a lower literal match with the reference text and thus a decrease in the ROUGE-L metric.

We further explored the impact of the LDM and VDM blocks in high-level tasks on the model results. As shown in Table 3, Baseline’ is the model with only high-level tasks removed from the complete three-layer structure. When using the LDM module alone, the BLEU-4 index increased by 0.007, and the ROUGE-L index increased by 0.038. This indicates that after the model achieved deconfounding in the language modality, its long-sentence matching ability improved, and the sentence structure became more reasonable. When using the VDM module alone, the BLEU-1 index increased by 0.019, and the CIDEr index increased by 0.025. This shows that after the model achieved deconfounding in the visual modality, its short-sentence matching ability and the ability to describe specific diseases improved. This may be related to the local key features added during the visual deconfounding process.

4.3.3. Case Study.

We conducted a case study on MIMIC-CXR and presented qualitative examples in Figure 4 to illustrate the excellent report generation ability of our method. **The red parts represent the incorrect or missing content in the BASELINE.** The green and blue parts in HTSC-CIF represent two types of logical associations: **1. Connectives, that is, context logic (green); 2. Logic of disease description (blue).** As can be seen from the results, we achieved good results in both context logic and the logic of disease description.

Dataset	Method	Metrics						
		B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
MIMIC-CXR	R2GEN [5]	0.353	0.218	0.145	0.103	0.128	0.267	-
	M2KT [46]	0.386	0.237	0.157	0.111	-	0.274	0.111
	KiUT [14]	0.393	0.243	0.159	0.113	0.16	0.285	-
	MET [37]	0.386	0.25	0.169	0.124	0.152	0.291	0.362
	DCL [19]	-	-	-	0.109	0.150	0.284	0.281
	RGRG [27]	0.373	0.249	0.175	0.126	0.168	0.264	0.495
	MAN [26]	0.396	0.244	0.162	0.115	0.151	0.274	-
	CMN [6]	0.353	0.218	0.148	0.106	0.142	0.278	-
	AlignTransformer [48]	0.378	0.235	0.156	0.112	0.158	0.283	-
	XrayGPT(7B) [29]	0.128	0.045	0.014	0.004	0.079	0.111	-
	Med-PaLM(12B) [30]	0.309	-	-	0.104	-	0.262	0.234
	Med-PaLM(562B) [30]	0.317	-	-	0.115	-	-	0.252
	R2GenGPT [38]	0.411	0.267	0.186	0.134	0.16	0.297	0.269
	Med-LMM [24]	-	-	-	0.128	0.161	0.289	0.265
	PromptMRG [16]	0.398	-	-	0.112	0.157	0.268	-
	VLCI [4]	0.400	0.245	0.165	0.119	0.150	0.280	0.190
	Ours	0.413	<u>0.264</u>	0.193	0.139	0.177	0.305	<u>0.386</u>
IU-Xray	R2GEN [5]	0.47	0.304	0.219	0.165	0.187	0.371	-
	M2KT [46]	0.497	0.319	0.23	0.174	-	0.399	0.407
	KiUT [14]	0.525	0.360	0.251	0.185	0.242	0.409	-
	MET [37]	0.483	0.322	0.228	0.172	0.192	0.38	0.435
	DCL [19]	-	-	-	0.163	0.193	0.383	0.586
	MAN [26]	0.501	0.328	0.23	0.170	0.213	0.386	-
	CMN [6]	0.475	0.309	0.222	0.170	0.191	0.375	-
	AlignTransformer [48]	0.484	0.313	0.225	0.173	0.204	0.379	-
	R2GenGPT [38]	0.488	0.316	0.228	0.173	0.211	0.377	0.438
	Med-LMM [24]	-	-	-	0.168	0.209	0.381	0.427
	VLCI [4]	0.505	0.334	0.245	0.189	0.204	0.397	0.456
	Ours	0.527	<u>0.356</u>	0.257	0.192	0.253	0.414	<u>0.538</u>

Table 1. The comparative results of metrics on the MIMIC-CXR and IU-Xray datasets are as follows. The research results are cited from the original papers. The best performance is marked in bold, and the second-best is underlined.

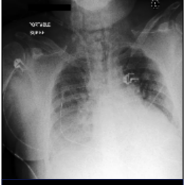
 MIMIC-CXR	Ground Truth Comparison is made to previous study from _____. Median sternotomy wires are again seen and unchanged. The heart size is within normal limits. There are densities at the bases suggestive of subsegmental atelectasis at the lung bases. This may represent early consolidation, however. The upper lung fields are clear. There are no signs of pulmonary edema. Bony structures are grossly intact.	BASELINE Unchanged median sternotomy wires are seen. Normal heart size is noted. Lung bases show densities possibly indicating subsegmental atelectasis (The inference of "early consolidation" in GT is missing). Upper lungs appear clear and pulmonary edema is observed (In GT, it is "no signs of"). Bony structures maintain gross integrity.	HTSC-CIF Image shows persistently unchanged median sternotomy wires. The cardiac silhouette remains normal with no pulmonary edema. Lung base densities suggest subsegmental atelectasis (Firstly, the densities at the lung base suggest subsegmental atelectasis, and further inference indicates early consolidation), though possibly representing early consolidation, while the upper lungs remain clear. Bony structures are grossly intact without acute fractures or lesions.
 IU-Xray	Ground Truth The trachea is midline. The cardiomeastinal silhouette is normal. The superior thoracic spine is again noted, unchanged from prior. Lucent pulmonary parenchyma is consistent appearance with emphysema and appears unchanged from prior examinations. No evidence of pneumothorax. No focal airspace disease or pleural effusion. Vague density in the medial right lung apex most XXXX representing overlying shadows of bony structures, which is stable.	BASELINE The trachea is midline and cardiomeastinal silhouette is normal. Pulmonary lucency suggests early emphysema, similar to prior. Mild pneumothorax is seen(The inference in GT is "No evidence of pneumothorax") without significant pleural effusion. Lung fields are clear without focal airspace disease. A subtle left(In GT, it is "right") apical density, possibly early infiltrate, remains stable. The upper thoracic spine shows no interval change.	HTSC-CIF The trachea is midline with a normal cardiomeastinal silhouette(The logic is manifested in that when the trachea is in the middle position, there will generally be a normal cardiac mediastinal contour). The pulmonary parenchyma demonstrates lucency consistent with emphysema (The logic is demonstrated by inferring emphysema based on abnormal transparency), unchanged compared to prior examinations. No pneumothorax, focal airspace disease, or pleural effusion is identified. A subtle opacity in the medial right apical region, most compatible with overlying bony structure shadows, remains stable. The superior thoracic spine shows no interval change from previous imaging.

Figure 4. Qualitative examples of HTSC-CIF on MIMIC-CXR.

Model			Metrics						
			B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
Baseline			0.304	0.187	0.126	0.096	0.124	0.270	0.293
Low	Mid	High							
-	-	✓	0.374	0.238	0.163	0.114	0.136	0.318	0.322
-	✓	✓	0.398	0.245	0.165	0.119	0.150	0.280	0.357
✓	-	✓	0.402	0.252	0.179	0.118	0.163	0.297	0.361
✓	✓	-	0.363	0.237	0.166	0.116	0.144	0.271	0.317
HTSC-CIF			0.413	0.264	0.193	0.139	0.177	0.305	0.386

Table 2. Exploring the impact of different levels of tasks on the model.

Model		Metrics						
		B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr
Baseline		0.363	0.237	0.166	0.116	0.144	0.271	0.317
LDM	VDM							
✓	-	0.367	0.239	0.167	0.123	0.146	0.309	0.335
-	✓	0.382	0.246	0.172	0.115	0.145	0.301	0.342
HTSC-CIF		0.413	0.264	0.193	0.139	0.177	0.305	0.386

Table 3. Exploring the influence of LDM and VDM modules in high-level tasks.

5. Conclusion

We proposed a novel Hierarchical Task Structure-Based Cross-Modal Causal Intervention Framework (HTSC-CIF) for MRG. This was the first time that the three challenging tasks of knowledge enhancement, cross-modal information alignment, and spurious correlation elimination in MRG were addressed in a hierarchical manner. Experiments showed that HTSC-CIF achieved SOTA performance on a medical image report benchmark dataset.

References

- [1] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision-language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022. 2
- [2] Dehua Chen, Hongjin Zhao, Jianrong He, Qiao Pan, and Weiliang Zhao. An causal xai diagnostic model for breast cancer based on mammography reports. In *2021 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 3341–3349. IEEE, 2021. 2
- [3] Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Shen Zhao, Guanbin Li, Cheng-Lin Liu, and Liang Lin. Cross-modal causal intervention for medical report generation. *arXiv preprint arXiv:2303.09117*, 2023. 2, 3, 4
- [4] Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Guanbin Li, Cheng-Lin Liu, and Liang Lin. Cross-modal causal representation learning for radiology report generation. *IEEE Transactions on Image Processing*, 34:2970–2985, 2025. 7, 8
- [5] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020. 8
- [6] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online, 2021. Association for Computational Linguistics. 8
- [7] Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*, 2022. 3
- [8] Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21361–21371, 2023. 3
- [9] Louke Delrue, Robert Gosselin, Bart Ilsen, An Van Landeghem, Johan de Mey, and Philippe Duyck. Difficulties in the interpretation of chest radiography. *Comparative interpretation of CT and standard radiography of the chest*, pages 27–49, 2011. 1
- [10] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 6
- [11] Iryna Hartsock and Ghulam Rasool. Vision-language models for medical report generation and visual question answering: A review. *arXiv preprint arXiv:2403.02469*, 2024. 2
- [12] Ju He, Jie-Neng Chen, Shuai Liu, Adam Kortylewski, Cheng Yang, Yutong Bai, and Changhu Wang. Transfg: A transformer architecture for fine-grained recognition. In *Proceedings of the AAAI conference on artificial intelligence*, pages 852–860, 2022. 6
- [13] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 4
- [14] Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818, 2023. 2, 8
- [15] Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*, 2021. 6
- [16] Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2607–2615, 2024. 8
- [17] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 6

- [18] Charles Jones, Daniel C Castro, Fabio De Sousa Ribeiro, Ozan Oktay, Melissa McCradden, and Ben Glocker. A causal perspective on dataset bias in machine learning for medical imaging. *Nature Machine Intelligence*, 6(2):138–146, 2024. 3
- [19] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343, 2023. 2, 8
- [20] Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2863–2874, 2023. 3
- [21] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762, 2021. 2
- [22] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Yuexian Zou, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*, 2021. 2
- [23] Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, 34:16266–16279, 2021. 3
- [24] Rui Liu, Mingjie Li, Shen Zhao, Ling Chen, Xiaojun Chang, and Lina Yao. In-context learning for zero-shot medical report generation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8721–8730, 2024. 8
- [25] Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, 2022. 2
- [26] Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4776–4783, 2024. 8
- [27] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023. 8
- [28] Yitian Tao, Liyan Ma, Jing Yu, and Han Zhang. Memory-based cross-modal semantic alignment network for radiology report generation. *IEEE Journal of Biomedical and Health Informatics*, 2024. 3
- [29] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023. 8
- [30] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024. 8
- [31] Athanasios Vlontzos, Daniel Rueckert, and Bernhard Kainz. A review of causality for learning algorithms in medical image analysis. *arXiv preprint arXiv:2206.05498*, 2022. 3
- [32] David Wadden, Ulme Wennberg, Yi Luan, and Hananeh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*, 2019. 7
- [33] Jun Wang, Abhir Bhalerao, and Yulan He. Cross-modal prototype driven network for radiology report generation. In *European Conference on Computer Vision*, pages 563–579. Springer, 2022. 3
- [34] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021. 6
- [35] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. 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*, pages 9049–9058, 2018. 2
- [36] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 4
- [37] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567, 2023. 8
- [38] Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033, 2023. 8
- [39] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21372–21383, 2023. 2
- [40] Xing Wu, Shaoqi Peng, Jingwen Li, Jian Zhang, Qun Sun, Weimin Li, Quan Qian, Yue Liu, and Yike Guo. Causal inference in the medical domain: a survey. *Applied Intelligence*, pages 1–24, 2024. 3
- [41] Xiaozheng Xie, Jianwei Niu, Xuefeng Liu, Zhengsu Chen, Shaojie Tang, and Shui Yu. A survey on incorporating domain knowledge into deep learning for medical image analysis. *Medical Image Analysis*, 69:101985, 2021. 3
- [42] Dexuan Xu, Huashi Zhu, Yu Huang, Zhi Jin, Weiping Ding, Hang Li, and Menglong Ran. Vision-knowledge fusion model for multi-domain medical report generation. *Information Fusion*, 97:101817, 2023. 3
- [43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 5

- [44] Sixing Yan, William K Cheung, Keith Chiu, Terence M Tong, Ka Chun Cheung, and Simon See. Attributed abnormality graph embedding for clinically accurate x-ray report generation. *IEEE Transactions on Medical Imaging*, 42(8): 2211–2222, 2023. [3](#)
- [45] Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510, 2022. [3](#)
- [46] Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798, 2023. [3](#), [8](#)
- [47] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11): 12996–13010, 2021. [4](#)
- [48] Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 72–82. Springer, 2021. [8](#)
- [49] Ke Yu, Shantanu Ghosh, Zhexiong Liu, Christopher Deible, and Kayhan Batmanghelich. Anatomy-guided weakly-supervised abnormality localization in chest x-rays. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 658–668. Springer, 2022. [7](#)
- [50] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 721–729. Springer, 2019. [3](#)
- [51] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12910–12917, 2020. [3](#)