Contrastive Learning with Counterfactual Explanations for Radiology Report Generation

Mingjie Li[1], Haokun Lin[2], Liang Qiu[1], Xiaodan Liang[2⋆], Ling Chen[3], Abdulmotaleb Elsaddik[2], and Xiaojun Chang[4,2]

[1] Stanford University, Palo Alto CA 94305 USA
[2] MBZUAI, Abu Dhabi, UAE
[3] University of Technology Sydney, Ultimo NSW 2007 Australia
[4] School of Information Science and Technology, University of Science and Technology of China
{lmj695,qiuliang}@stanford.edu, ling.chen@uts.edu.au
{haokun.lin,xiaodan.liang,a.elsaddik,xiaojun.chang}@mbzuai.ac.ae

Abstract. Due to the common content of anatomy, radiology images with their corresponding reports exhibit high similarity. Such inherent data bias can predispose automatic report generation models to learn entangled and spurious representations resulting in misdiagnostic reports. To tackle these, we propose a novel **Co**unter**F**actual **E**xplanations-based framework (CoFE) for radiology report generation. Counterfactual explanations serve as a potent tool for understanding how decisions made by algorithms can be changed by asking "what if" scenarios. By leveraging this concept, CoFE can learn non-spurious visual representations by contrasting the representations between factual and counterfactual images. Specifically, we derive counterfactual images by swapping a patch between positive and negative samples until a predicted diagnosis shift occurs. Here, positive and negative samples are the most semantically similar but have different diagnosis labels. Additionally, CoFE employs a learnable prompt to efficiently fine-tune the pre-trained large language model, encapsulating both factual and counterfactual content to provide a more generalizable prompt representation. Extensive experiments on two benchmarks demonstrate that leveraging the counterfactual explanations enables CoFE to generate semantically coherent and factually complete reports and outperform in terms of language generation and clinical efficacy metrics.

Keywords: Medical report generation · Counterfactual explanation · Contrastive learning

1 Introduction

Automatically generating reports can reduce the load on radiologists and potentially increase the accuracy and consistency of interpretations. This is achieved

⋆ Corresponding Author. Code is available at: https://github.com/mlii0117/CoFE

by translating intricate radiology images into semantically coherent and clinically reliable free texts. However, in comparison to generic captioning tasks, Radiology Report Generation (RRG) presents a significant challenge, often yielding unsatisfactory performance when employing direct captioning methods [30, 42] in the field of radiology. The difficulty arises due to the severe data bias within the limited image-report pair data available, a challenge that has been extensively acknowledged and discussed [4, 16, 23, 27, 36, 45].

Given the shared anatomical content, radiology images tend to display significant similarity to one another, with abnormal or lesioned areas typically occupying minor portions of the images [24, 44]. This similarity also extends to the accompanying reports, where several sentences often describe normal tissues. However, the clinical usefulness of radiology reports hinges on the accurate depiction of abnormalities. This intrinsic data bias tends to lead models to learn spurious and intertwined visual features, resulting in the generation of inaccurate diagnostic reports. To mitigate data bias, various successful concepts have been proposed by existing methods to enhance learning representations, such as employing contrastive learning [23, 28], incorporating medical knowledge [27, 49], and implementing relational memory [4] etc.

Recently, Tanida *et al.* [36] achieved promising performances by detecting abnormal regions using a pre-trained detector with pseudo labels. They then utilized these features to guide a pre-trained large language model (LLM) in generating reports. Identifying critical regions that cover abnormalities or lesions not only enhances non-spurious visual representation but also improves the explainability of RRG models. Ideally, the method would employ golden annotations to train a lesion detector capable of accurately localizing abnormal regions or lesions. However, existing RRG benchmarks lack such annotations. Relying on weakly supervised signals from pseudo labels [36, 46] can result in misalignment. Furthermore, the limited size of available medical data may prevent the full unleashing of the potential of LLMs.

To address these challenges, we introduce a novel concept: counterfactual explanation (CE). The concept of CE [12, 43] has surfaced in machine learning as an insightful tool to comprehend how models' decisions can be changed. CE offers a hypothetical alternative to the observed data, allowing for the assessment and comprehension of models through 'what if' scenarios. This technique has been integrated into diagnostic models to not only improve diagnostic accuracy but also enhance explainability [5, 37]. For example, Tanyel *et al.* [37] proposed a CE to identify the minimal feature change and effectively demonstrated which features are more informative in differentiating two tumor types from MRI brains. Inspired by this, we aim to make this progress further interactive and explainable by explaining the global feature change in specific local regions. In particular, we propose a CE as ***'what if we exchange the patch between two images, will the diagnosis shift?'*** to identify critical regions within images that may cover abnormalities or lesions, providing insights into the diagnosis process. For instance, as illustrated in Figure.1, we generate a counterfactual image by iteratively swapping a patch between semantically similar images with

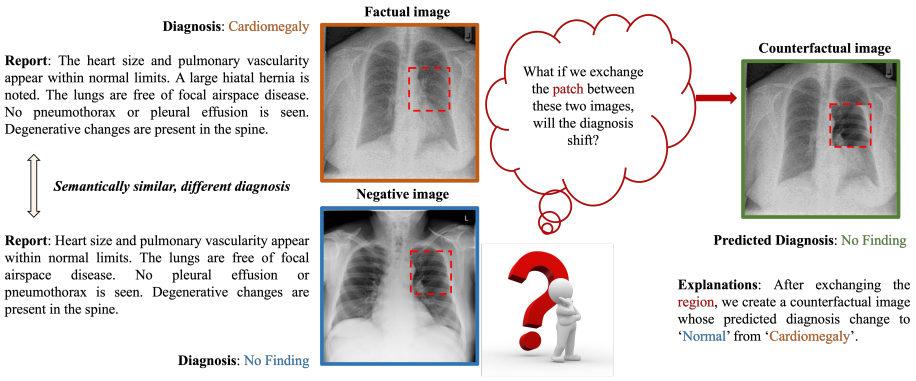


Fig. 1: A conceptual overview of our proposed counterfactual explanations is presented. Such CEs help to construct a counterfactual image by iteratively exchanging a patch between factual (positive) and negative images until the predicted diagnosis shift occurs. In this instance, the box in red covering the heart is identified as the critical region that causes the diagnosis shift.

different diagnosis labels until a shift in predicted diagnosis is achieved. Due to aforementioned similarities, exchanging a patch in the same position between two radiology images, particularly those that are semantically similar but carry different labeled diagnoses, does not disrupt the anatomical content. Notably, previous methods only integrate the CE into the decision-making process, lacking the ability to convey factual or counterfactual information effectively. In contrast, we translate our CE into a prompt that can present the key concept of CE and encapsulate the factual and counterfactual content. This prompt can yield more comprehensive instructions to LLMs and facilitate the elicitation of their knowledge.

In this paper, we propose a **Co**unter**F**actual **E**xplanations-based framework (CoFE) for radiology report generation. CoFE is capable of learning non-spurious visual representations and effectively leverage the capabilities of LLMs. First, we introduce a novel type of CEs for RRG tasks and propose a counterfactual generation process through contrastive learning, which constructs a counterfactual image and a learnable prompt. Specifically, we adopt a negative sampling strategy to discover the most semantically similar negative sample from the data bank, based on text similarity and diagnosis label. By iteratively exchanging patches between factual (positive) and negative samples until a predicted diagnosis change occurs, we pinpoint the critical region and create a counterfactual image. We then employ contrastive learning within a joint optimization framework to differentiate representations between factual and counterfactual samples, enabling the model to learn non-spurious visual representations. Subsequently, we employ a pretrained LLM, GPT-2 Medium [34], as a decoder to generate reports. To fine-tune the LLM efficiently, we propose a learnable prompt that encapsulates both factual and counterfactual content. This prompt can elicit the embedded knowledge within the LLM, which is helpful to generate semantically coherent and factually complete reports.

We evaluate our proposed method on two benchmarks, IU-Xray [6] and MIMIC-CXR [18] respectively. Extensive experiments demonstrate that our approach can outperform previous competitive methods in metrics that measure descriptive accuracy and clinical correctness. It indicates that leveraging CEs to learn non-spurious representations and prompt the generation process can improve the quality of predicted reports.

2 Related Work

2.1 Medical Report Generation

The pursuit of automating medical report generation through machine learning aims to alleviate the workload on radiologists. Numerous effective concepts exist to learn non-spurious representations to mitigate inherent data bias. Relation memory [3, 4] can prompt enhancement by boosting interaction efficiency of cross-modal memory network. Integrating medical knowledge is another solution, researchers utilize graph structural data [23, 48] or medical tags [17, 22] to incorporate prior knowledge into image encoding. Additional models [47, 50] also enhance performance by integrating knowledge information, with strategies including multi-modal semantic alignment and multi-label classification pretraining. To identify the abnormalities, PPKED [27] employs a unique architecture to mimic human learning processes. Tanida *et al.* [36] utilize a lesion detector pre-trained by pseudo labels to attain the non-spurious features and lead a pretrained GPT-2 [34]. Due to the lack of annotations, weakly supervised signals from pseudo labels may lead to misalignment. Although, large pretrained models showcase the adaptability in learning medical representations [31], the scarcity of data may limit the potential of LLMs. In this paper, our method concentrates on learning non-spurious representations by identifying critical regions and employing a robust prompt to fine-tune the LLM efficiently.

2.2 Counterfactual Explanations Reasoning

The advent of counterfactual explanations (CEs) construction has driven significant innovations, particularly in computer vision applications, enhancing both accuracy and interpretability. CEs have the potential to relieve existing methodologies from the reliance on extensive training data and meticulous annotations, by asking 'what if' scenario to explore self-supervised signals. Fang *et al.* [8] and Kim *et al.* [19] have introduced systems and frameworks, such as Counterfactual Generative Networks (CGN), designed to augment interpretability and resilience to CEs inputs without compromising accuracy. Similarly, CPL [12] proficiently generates counterfactual features and has exhibited remarkable efficacy in tasks like image-text matching and visual question answering. Ji *et al.* [15] specifically target video relationship detection, constructing CEs to elucidate their influence on factual scenario predictions. Further, the studies by Yang *et al.* [51] on PubMedQA highlight the crucial role of CEs, generated via ChatGPT, in learning

causal features in counterfactual classifiers, demonstrating the versatility and broad applicability of counterfactual methods across various domains. In this paper, we employ CEs to enhance the RRG models, especially where acquiring a substantial amount of golden annotations is prohibitively expensive.

2.3 Prompt Tuning

Prompt tuning is a method in natural language processing (NLP) used to efficiently modify the behavior of a pre-trained LLM, based on specific prompts or trigger phrases. This approach involves fine-tuning the model on a set of prompts or queries and their corresponding responses, allowing the model to respond more accurately or appropriately to those or similar prompts [35, 54]. For example, Guo *et al.* [11] applied Q-Learning to optimize soft prompts, while PTuning v2 [29] demonstrated that continuous prompt tuning could match the performance of fine-tuning in various scenarios. This technique has also garnered significant interest in the field of computer vision. CoOp [33] introduced a strategy for continuous prompt optimization to negate the need for prompt design, and CoCoOp [53] expanded upon this by learning an instance-conditional network to generate a unique input-conditional token for each image. Fischer *et al.* [9] also prove the adaptability of prompt tuning in medical image segmentation tasks. However, the reliance on empirical risk minimization presents challenges, necessitating advancements to avoid spurious representations. In this paper, we aim to propose a generalizable prompt incorporating factual and counterfactual content to efficiently refine the medical LLMs.

3 Methodology

In this section, we introduce the detailed implementations of our proposed **Co**unter**F**actual **E**xplanations-based framework (CoFE). As shown in Fig.2, our CoFE mainly consists of two uni-modal encoders, one cross-modal encoder, a language decoder, and a counterfactual generation module with four training objectives. We first introduce the backbone of CoFE and then describe the counterfactual generation process in detail.

3.1 Set Up

In this work, we aim to integrate counterfactual explanations into report generation models to learn non-spurious visual representations and efficiently generate high-quality reports. Radiology report generation tasks require a model to translate a complex radiology image I into a generic report $T = \{y_1, y_2, \ldots, y_n\}$. We denote the target report by $\hat{T} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{\hat{n}}\}$. n and $\hat{n}$ represent the number of tokens in a report. In addition to corresponding reports, we also utilize the diagnosis label C for each examination in this work. Since not all existing benchmarks provide such annotations, we use the CheXpert labeling tool [13] to label ground truth reports with 14 different medical terminologies. Notably, we assign

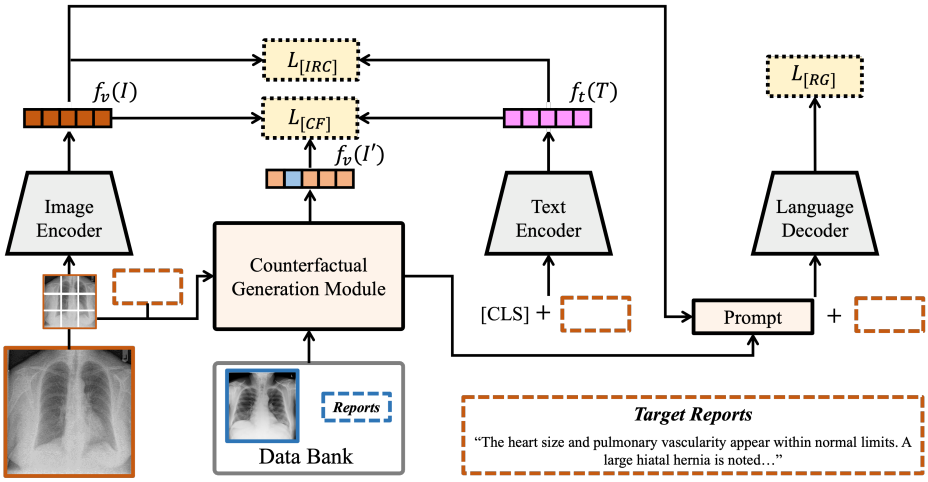


Fig. 2: Illustration of our proposed **Co**unter**F**actual **E**xplanations-based framework (CoFE). CoFE consists of two unimodal encoders, one cross-modal encoder, one language decoder, and our proposed counterfactual generation module that can construct a counterfactual image and a learnable prompt, respectively. The entire framework is trained through joint optimization, mainly employing contrastive learning paradigms for radiology report generation.

the label "No Finding" when CheXPert does not extract any terminologies. For instances with multiple labels, we adopt a strategy during training where only one label is randomly retained for each case.

Automatic report generation systems are typically based on encoder-decoder frameworks. The encoder generally aims to convert the given image I into dense visual vectors $f_v(I)$. The decoder is usually a sequence processing network, which translates $f_v(I)$ to a report T. To generate the CE and calculate the image report contrastive loss, we involve an additional text encoder following the BLIP [21] architecture, drawing inspiration from the successful concepts found in [23, 28].

3.2 Encoders

Image encoder. Different from prior works employing CNNs, we use a pre-trained ViT [7]-S as the image encoder $f_v(\cdot)$. ViT enables finer semantic feature extraction by dividing images into more patches, whose resolution is 16×16, compared to conventional CNNs 7×7. A `[CLS]` token is also prepended before input to the encoder layers. The encoder layer process, $\mathbf{f}_e(\cdot)$, is defined as:

$$\mathbf{f}_e(x) = \mathrm{LN}(\mathrm{FFN}(e_{attn}) + e_{attn}), \tag{1}$$

$$e_{attn} = \mathrm{LN}(\mathrm{MHA}(x) + x), \tag{2}$$

where FFN and LN represent Feed Forward Network [40] and Layer Normalization operation [1], respectively. MHA [40] (multi-head attention) splits attention

into n heads, with each head, $\mathrm{Att}(\cdot)$, defined as:

$$\mathrm{Att}(x) = \mathrm{softmax}(\frac{\mathbf{Q}^x(\mathbf{K}^x)^\top}{\sqrt{d}})\mathbf{V}^x. \tag{3}$$

with $d = 384$ being the embedding space dimension, and $\{\mathbf{Q}, \mathbf{K}^*, \mathbf{V}^*\}$ representing the corresponding *Query, Key, Value* vectors. The resulting output, the encoded visual vectors $\mathbf{f}_I$, will be used for report generation.

Text encoder. We employ a PubMedBERT [10], pre-trained with abstracts and full texts from PubMed[5], as our text encoder $f_t(\cdot)$. It extracts textual representations $f_t(T)$ from positive and negative reports, which will be utilized for calculating the image-report contrastive (IRC) loss to facilitate learning robust and generalizable medical visual and textual representations. We utilize momentum image and text encoders to extract positive and negative data representations in a batch. Then we first calculate the softmax-normalized image-to-report similarity $f_m^{\mathrm{i2t}}(I)$ and the report-to-image similarity $f_m^{\mathrm{t2i}}(T)$ for the image I and its paired report T by $f_m^{i2r}(I) = \frac{\exp s(I,T_m)/\tau}{\sum_{m=1}^{M} \exp s(I,T_m)/\tau}$, with τ as a learnable temperature parameter. The IRC loss can be written as:

$$\mathcal{L}_{\mathrm{IRC}} = \frac{1}{2}(\mathcal{L}_{\mathrm{ce}}(g^{t2i}(T), f^{t2i}(T)) + \mathcal{L}_{\mathrm{ce}}(g^{i2t}(I), f^{i2t}(I))). \tag{4}$$

where $g(\cdot)$ denotes the ground truth of similarity.

3.3 Decoding Process

Recognizing the advanced capabilities of Large Language Models (LLMs) in various language generation tasks, we utilize a GPT-2 Medium model, which is pre-trained on the PubMed dataset, as our language decoder. In contrast to R2GenGPT, which employs a frozen LLaMa-7B [38], our choice of GPT-2, with its 355 million parameters, allows for full fine-tuning. This adaptability enables GPT-2 to more effectively cater to the nuances of report generation tasks. GPT-2, an auto-regressive model leveraging self-attention, conditions each output token in a sequence on its previous tokens for report generation. The entire process can be represented as:

$$p(T|I) = \prod_{t=1}^{n} p(y_t|y_1, \ldots, y_{t-1}, I). \tag{5}$$

Here, y_t is the input token at time step t. The typical objective for report generation is minimizing cross-entropy loss between the predicted and ground truth token sequences. With ground truth report $\hat{R}$, all modules are optimized to maximize $p(\mathbf{y}|I)$ by minimizing:

$$\mathcal{L}_{\mathrm{RG}} = -\sum_{t=1}^{\hat{n}} \log p(\hat{y}_t|\hat{y}_1, \cdots, \hat{y}_{t-1}, I). \tag{6}$$

[5] https://pubmed.ncbi.nlm.nih.gov

3.4 Counterfactual Generation

In this section, we will explain how to generate counterfactual features, encompassing a counterfactual image and a learnable prompt in detail. Counterfactual images are pivotal, allowing the model to discern non-spurious features through contrasting representations between factual and counterfactual images. The learnable prompt encapsulates both factual and counterfactual contents and then efficiently refines the pre-trained LLM.

Negative sampling strategy. Counterfactual features combine features derived from both factual (positive) and negative data. We first propose a negative sampling strategy to select the negative data from a data bank. Such negative data should have different labels and are difficult to distinguish from factual data. To implement this, we first construct a data bank, denoted by D, containing candidate data, each instance d_i is annotated with {Image I_i, Report T_i, Label C_i}, maintaining the balanced distribution of diagnostic labels within the data bank. Next, we select the negative data from the data bank, as $d^- = \arg\max_i \text{BLEUScore}(T, T_i)$ and $C \neq C_i$. The BLEUScore function calculates the BLEU [32] score, setting the factual report as a reference and the negative data as a candidate. The so-selected $d^- = I^-, T^-, C^-$ is earmarked as negative data, exhibiting textual semantics similar to the original data but possessing distinct labels, emphasizing their inherent dissimilarity. The entire procedure is visually depicted in Fig. 3.

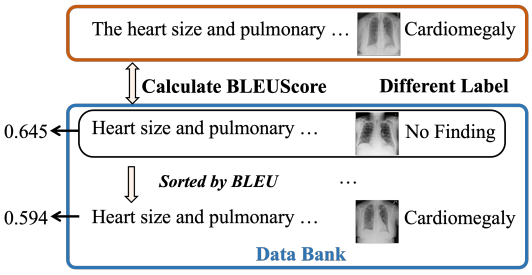


Fig. 3: Illustration of negative sampling strategy. The objective is to select a negative sample that is mostly similar in semantics but carries a different diagnostic label from the data bank.

Counterfactual image. After selecting the negative data from candidates, we proceed to generate counterfactual images combining factual and negative images, thereby enhancing non-spurious representations through contrastive learning. As shown in Fig.4, a factual image, I, is presented in the form of n patches: $I = p_1, p_2, ..., p_n$; its corresponding negative image is represented as $I^- = p_1^-, p_2^-, ..., p_n^-$. From the 1st to the n-th patch, each patch of the negative image replaces the patch of the factual image at the corresponding position. The modified image is denoted by $I' = (1-u) * I + u * I^-$, where u is a one-hot vector to present the index of the replaced patch. Subsequent to each replacement, the modified image is fed into a pre-trained and frozen discriminator composed of the image encoder and a Multilayer Perceptron (MLP) to predict the logits for the diagnostic label C'. The replacement process ceases once $C' \neq C$, culminating in the acquisition of the counterfactual image I'. This methodology

enables the identification of critical regions that prompt models to alter the predicted diagnosis. In essence, such regions contain pivotal information pertinent to the examination. It helps to mitigate inherent data bias and facilitates the model's focus on these critical regions, learning non-spurious and robust visual representations.

Learnable prompt. Another key component of our counterfactual features is a learnable prompt, designed to elicit knowledge and leash the potential of pre-trained LLMs. Frequently used prompts in caption tasks, such as "the caption is..." or "describe [visual tokens]", clarify the task but often lack comprehensive instructions. To rectify this deficiency, we embed both factual and counterfactual content within the learnable prompt to attain more generalizable representations. As suggested by [39], our prompt incorporates detailed instructions and is formulated by concatenating the factual visual tokens, factual label, counterfactual label, and the index of the patch with supplementary text. The training prompt is articulated as "The u patch of image contains critical features for diagnosing C. Generate a diagnostic report for the image by describing critical entities including tubes, pneumothorax, pleural effusion, lung opacity, cardiac silhouette, hilar enlargement, and mediastinum."

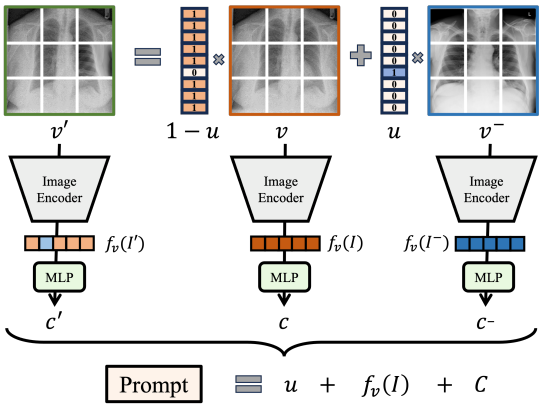


Fig. 4: Illustration of the counterfactual generation process, including a counterfactual image and a learnable prompt.

3.5 Joint optimization

In addition to the image-report contrastive loss, and report generation loss, we introduce a novel contrastive loss aimed at amplifying the proficiency of visual representation learning. Specifically, the factual image feature $f_v(I)$, text feature $f_t(T)$, and counterfactual image feature $f_v(I')$ are employed to compute the counterfactual loss, $\mathbf{L}_{cf}$, thereby extending the divergence between the counterfactual features and the features of the original data. This can be represented as:

$$\mathcal{L}_{CF} = -\log \frac{e^{\frac{f_v(I)@f_t(T)}{\tau}}}{e^{\frac{f_v(I)@f_t(T)}{\tau}} + e^{\frac{f_v(I')@f_t(T)}{\tau}}} \tag{7}$$

Here, @ denotes cosine similarity. The total training loss is written as:

$$\mathcal{L} = \mathcal{L}_{IRC} + \lambda_{rg}\mathcal{L}_{RG} + \lambda_{cf}\mathcal{L}_{CF} \tag{8}$$

where λ_{cf} and λ_{rg} denote the loss weights, we assign a value of 1.2 to λ_{cf} and 1.5 to λ_{rg} based on performance on the validation set.

3.6 Inference

Our inference methodology is streamlined, as the counterfactual generation module is operational exclusively during the training phase, thereby enhancing the learning of both visual and textual representations. Given an image, the hidden embeddings $f_V(I)$ are seamlessly concatenated with the prompt and fed into the LLM to generate diagnostic reports. In a similar vein, the prompt is refined to "Generate a diagnostic report for the image by describing critical entities including tubes, pneumothorax, pleural effusion, lung opacity, cardiac silhouette, hilar enlargement, and mediastinum."

4 Experiments

4.1 Datasets, Evaluation Metrics and Settings

Datasets. We validate the efficacy of our proposed CoFE using the **IU-Xray** [6] and **MIMIC-CXR** [18] benchmarks. The settings adopted by [4] are utilized to uniformly split and preprocess the datasets and reports, ensuring a fair comparison. IU-Xray [6], a prevalent benchmark for evaluating RRG systems, comprises 3,955 reports and 7,470 images. After excluding cases with only one image as per [4, 20], 2069/296/590 cases are allocated for training/validation/testing respectively. We utilize CheXPert to extract terminologies from reports and assign labels to each examination. MIMIC-CXR [18], the most extensive radiology dataset publicly available, includes 368,960 images and 222,758 reports. It has officially segmented subsets and has spurred the development of structurally explorative child datasets like RadGraph [14].

Metrics. We employ two types of metrics to evaluate the quality of our predicted reports. First, **natural language generation** (NLG) are employed to assess the descriptive precision of the predicted reports, with CIDEr [41] and BLEU [32] being primary. BLEU is primarily designed for machine translation, evaluating word n-gram overlap between reference and candidate, repeating frequent sentences can also achieve high scores. Conversely, CIDEr, developed for captioning systems, rewards topic terms and penalizes frequent ones, thus is more fitting for evaluating reports in RRG tasks. Additionally, ROUGE-L [25] and METEOR [2] are also considered for comprehensive comparison. Lastly, **clinical efficacy** metrics, a more recent innovation, ascertain the clinical accuracy of reports by using the CheXPert labeling tool to annotate predicted reports. Subsequent classification measurements like F1-Score, Precision, and Recall assess the aptness of the generated reports in describing abnormalities.

Table 1: The performance in NLG metrics of our proposed method compared to other competitive methods on the IU-Xray and MIMIC-CXR datasets. The highest figures in each column are highlighted in bold.

IU-Xray					MIMIC-CXR				
Methods	CIDEr	BLEU-4	ROUGE-L	METEOR	Methods	CIDEr	BLEU-4	ROUGE-L	METEOR
R2Gen	0.398	0.165	0.371	0.187	R2Gen	0.253	0.103	0.277	0.142
KERP	0.280	0.162	0.339	-	CMN	-	0.106	0.278	0.142
HRGP	0.343	0.151	0.322	-	TopDown	0.073	0.092	0.267	0.129
MKG	0.304	0.147	0.367	-	PPKED	0.237	0.106	0.284	0.149
PPKED	0.351	0.168	0.376	0.190	RGRG	**0.495**	**0.126**	0.264	0.168
MGSK	0.382	**0.178**	0.381	-	MGSK	0.203	0.115	0.284	-
CMCL	-	0.162	0.378	0.186	CMCL	-	0.097	0.281	0.133
DCL	0.586	0.163	0.383	0.193	DCL	0.281	0.109	0.284	0.150
CoFE	**0.731**	0.175	**0.438**	**0.202**	**CoFE**	0.453	0.125	**0.304**	**0.176**

Experimental settings. For both datasets, we only utilize the front view examinations. We first pre-train the ViT-S for 10 epochs using diagnosis labels. Given the distinct domain difference between medical and general texts, a pretrained PubMedBert [10] is utilized as both a tokenizer and a text encoder. The training is conducted on 4 NVIDIA 2080 Ti GPUs, spanning 50 epochs with batch sizes of 8. The model checkpoint achieving the highest CIEDr metric is selected for testing. An Adam optimizer, with a learning rate of 1e-4 and a weight decay of 0.02, is applied. The learning rate drops 10 every 2 epochs and stops at 1e-6. We set the size of data bank to 1,380. Note that all encoded vectors are projected by a linear transformation layer into a dimension of $d = 384$.

4.2 Main Results

Descriptive Accuracy. We compare our CoFE with several competitive RRG methods on two benchmarks. R2Gen [4] and CMN [3] are two widely-used baseline models implementing relation memory. KERP [20], PPKED [27], MKG [52] and MGSK [48] are proposed to integrate medical knowledge with typical RRG backbones. CMCL [26] and DCL [23] employ contrastive learning to further improve performance. As presented in Table.1, our method notably outperforms all competing approaches, attaining the highest figures across almost all the metrics, with a CIDEr score of 0.731 and a BLEU-4 score of 0.175 on IU-xray. Similarly, our method demonstrates competitive performance on the MIMIC-CXR dataset, achieving the highest ROUGE-L score of 0.304 and METEOR score of 0.176. These results showcase the superior capability of our method in generating matched and semantically similar reports.

Clinical Correctness. We also evaluate our method by Clinical Efficacy (CE) metrics on the MIMIC-CXR dataset to evaluate the clinical correctness of our predicted reports. In Table. 2, we compare the performance against several baseline models, DCL, R2Gen, and MKSG, respectively. Most notably, our CoFE achieves the SOTA performance across all the clinical efficacy metrics, with

a Precision of 0.489, Recall of 0.370, and F1-score of 0.405. This performance-boosting underscores the effectiveness of integrating counterfactual explanations, enabling the model to generate more clinically correct and relevant reports.

4.3 Analysis

In this section, we conduct ablation studies and a case study on IU-Xray and MIMIC-CXR datasets to investigate the proficiency of each key component in CoFE. Specifically, Table. 3 presents the quantitative analysis of CoFE on IU-Xray measuring descriptive accuracy. In addition, clinical correctness evaluation is reported in Table. 2. We employ a BLIP without the cross-modal encoder as our base model. We found that abandoning this module can save computation sources and achieve similar performances.

Table 2: The comparison of the clinical efficacy metrics on the MIMIC-CXR dataset.

Methods	Precision	Recall	F1-score
DCL	0.471	0.352	0.373
R2Gen	0.333	0.273	0.276
MKSG	0.458	0.348	0.371
Base	0.325	0.271	0.273
+ LLMs	0.396	0.323	0.317
+ prompt	0.463	0.355	0.366
+ $\mathcal{L}_{CF}$ (full)	**0.489**	**0.370**	**0.405**

Effect of pre-trained LLMs. Compared with the base model in setting (a), illustrated in Table 3, where we utilize a pre-trained PubMedBert and a 355M-parameter GPT-2 Meduium as the text encoder and language decoder, there is a significant enhancement in all metrics, with CIDEr improving from 0.363 to 0.510, emphasizing the impactful role of LLMs in enhancing the report generation performance. Specifically, PubMedBert can encode the reports into better textual representations, while GPT-2 has the capability to generate semantically coherent and logically consistent reports.

Non-spurious Representation Learning. The primary motivation for integrating counterfactual explanations is to enhance non-spurious visual representations by contrasting the representations between factual and counterfactual images. When comparing setting (c) to Setting (a) and the full model to setting (b), a significant performance boost is observable across all metrics. For instance, CIDEr elevates from 0.510 to 0.678, and from 0.698 to 0.731, respectively. Additionally, the final BLEU-4 metrics reach 0.175, achieving the SOTA performances. These notable elevations highlight the importance of non-spurious representation learning capabilities in radiology report generation tasks.

Effect of Prompt Tuning. To fully elicit pre-trained knowledge and unleash the potential of LLMs, we propose a learnable prompt that encapsulates both factual and counterfactual content to refine the LLMs. Observing setting (a) *vs* (b) and (c) *vs* the full model, it is evident that our proposed prompt can further augment performance, especially evident in ROUGE-L, which elevates from 0.355 to 0.373 and from 0.381 to 0.428, respectively. This increment underscores the effectiveness of our prompt in refining the model's natural language generation capability. Furthermore, as shown in Table.2, this prompt can also increase the clinical correctness of the predicted reports.

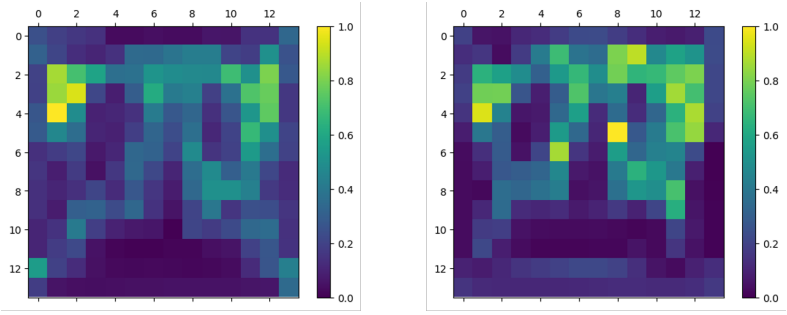
Fig. 5: Illustration of reports generated by R2Gen, DCL and our CoFE. The text in blue demonstrates the ground truth diagnosis labels. The red text represents the accurately matched abnormalities.

Negative Sampling Strategy. The key point to constructing a counterfactual image is selecting negative data which have different labels and are difficult to be distinguished from the factual data. To verify this, we employ a random sampling strategy in which candidate data are indiscriminately selected as the negative sample. The incorporation of this random sampling strategy in settings (c) and (d) results in a discernible degeneration in the model's capability to generate high-quality reports. For instance, the CIDEr metrics drop from 0.678 to 0.653, while Bleu-4 scores decrease to 0.156. This slight decline across almost all performance metrics elucidates the influential role of our negative sampling strategy in pinpointing the most suitable negative data.

Table 3: Quantitative analysis of our proposed method on the IU-Xray dataset. We employ a vanilla BLIP without loading pre-trained parameters as the base model.

Settings	LLMs	Prompt	$\mathcal{L}_{CF}$	Random Sampling	CIDEr	BLEU-4	ROUGE-L	METEOR
Base					0.363	0.132	0.248	0.154
(a)	✓				0.510	0.151	0.355	0.180
(b)	✓	✓			0.698	0.160	0.373	0.183
(c)	✓		✓		0.678	0.164	0.381	0.191
(d)	✓		✓	✓	0.653	0.156	0.392	0.186
(e)	✓	✓	✓	✓	0.706	0.166	0.407	0.196
CoFE	✓	✓	✓		**0.731**	**0.175**	**0.428**	**0.202**

Qualitative Analysis. In Figure.5, we present two samples from MIMIC-CXR and corresponding reports generated by R2Gen, DCL and our CoFE. R2Gen seems to lack specificity and detailed insights, providing a more generalized statement about the conditions and missing several key abnormalities mentioned in the ground truth, such as pulmonary nodules and pleural effusion. The DCL model is somewhat more aligned with the ground truth, acknowledging the unchanged appearance of the cardiac silhouette and the presence of extensive bilateral parenchymal opacities. However, it fails to mention the presence of pulmonary nodules and the pleural effusion in the right middle fissure specifically. In contrast, CoFE addresses pleural effusion, atelectasis, and the absence of pneu-



(a) IU-Xray (b) MIMIC-CXR

Fig. 6: Heatmaps that illustrate the frequency at which individual patches were replaced during the training process of two distinct medical imaging datasets: IU-Xray on the left and MIMIC-CXR on the right.

monia and pneumothorax, making it more in alignment with certain elements of the ground truth. These observations prove that our CoFE can generate factual complete and consistent reports.

Frequency of Replaced Patches: In Figure 6, we present two heatmaps that illustrate the frequency at which specific patches were replaced to construct counterfactual images during the training process on the IU-Xray and MIMIC-CXR datasets. The color gradient in each patch ranges from dark to light, with the lightest shade denoting the highest replacement frequency. The observed patterns across the heatmaps indicate that patches associated with critical anatomical regions, such as the heart, lungs, and air spaces, are replaced more frequently. This suggests that during training, these areas are particularly targeted as they likely contain critical diagnostic information (counterfactual contexts), underlining their clinical significance in disease detection or condition assessment. The distribution of frequency substantiates our methodology for constructing counterfactual images, which effectively identifies and emphasizes the most pertinent regions, thereby potentially enhancing the representation learning process.

5 Conclusion

In this paper, we present a novel framework, Counterfactual Explanations-based Framework (CoFE), designed for radiology report generation. To address the inherent data bias, we introduce a novel counterfactual concept, allowing CoFE to identify critical regions and construct a counterfactual image during training. By contrasting the representations between factual and counterfactual features, CoFE is adept at learning non-spurious visual representations. Subsequently, we summarize the counterfactual generation process into a learnable prompt, enabling the efficient fine-tuning of a pre-trained LLM. Experiments on two widely-recognized benchmarks verify the efficacy of our approach in generating factual, comprehensive, and coherent reports.

Acknowledge This work is supported by ARC DP210101347.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016) 6
2. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization. pp. 65–72 (2005) 10
3. Chen, Z., Shen, Y., Song, Y., Wan, X.: 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). pp. 5904–5914 (2021) 4, 11
4. Chen, Z., Song, Y., Chang, T., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (2020) 2, 4, 10, 11
5. Dai, X., Keane, M.T., Shalloo, L., Ruelle, E., Byrne, R.M.J.: Counterfactual explanations for prediction and diagnosis in XAI. In: Conitzer, V., Tasioulas, J., Scheutz, M., Calo, R., Mara, M., Zimmermann, A. (eds.) AIES '22: AAAI/ACM Conference on AI, Ethics, and Society, Oxford, United Kingdom, May 19 - 21, 2021. pp. 215–226. ACM (2022). https://doi.org/10.1145/3514094.3534144, https://doi.org/10.1145/3514094.3534144 2
6. 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) 4, 10
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020) 6
8. Fang, Z., Kong, S., Fowlkes, C., Yang, Y.: Modularized textual grounding for counterfactual resilience. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6378–6388 (2019) 4
9. Fischer, M., Bartler, A., Yang, B.: Prompt tuning for parameter-efficient medical image segmentation. CoRR **abs/2211.09233** (2022). https://doi.org/10.48550/arXiv.2211.09233, https://doi.org/10.48550/arXiv.2211.09233 5
10. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing (2020) 7, 11
11. Guo, H., Tan, B., Liu, Z., Xing, E.P., Hu, Z.: Efficient (soft) q-learning for text generation with limited good data. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022. pp. 6969–6991. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.findings-emnlp.518, https://doi.org/10.18653/v1/2022.findings-emnlp.518 5
12. He, X., Yang, D., Feng, W., Fu, T.J., Akula, A., Jampani, V., Narayana, P., Basu, S., Wang, W.Y., Wang, X.: Cpl: Counterfactual prompt learning for vision and language models. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3407–3418 (2022) 2, 4

13. 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) 5
14. Jain, S., Agrawal, A., Saporta, A., Truong, S., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., Langlotz, C., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021) 10
15. Ji, X., Chen, J., Wu, X.: Counterfactual inference for visual relationship detection in videos. In: 2023 IEEE International Conference on Multimedia and Expo (ICME). pp. 162–167. IEEE (2023) 4
16. Jin, H., Che, H., Lin, Y., Chen, H.: Promptmrg: Diagnosis-driven prompts for medical report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 2607–2615. No. 3 (2024) 2
17. Jing, B., Xie, P., Xing, E.: On the automatic generation of medical imaging reports. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2577–2586 (2018) 4
18. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., ying Deng, C., Mark, R.G., Horng, S.: Mimic-cxr: A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019) 4, 10
19. Kim, J., Kim, M., Ro, Y.M.: Interpretation of lesional detection via counterfactual generation. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 96–100. IEEE (2021) 4
20. Li, C.Y., Liang, X., Hu, Z., Xing, E.P.: Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 6666–6673 (2019) 10, 11
21. Li, J., Li, D., Xiong, C., Hoi, S.C.H.: BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvári, C., Niu, G., Sabato, S. (eds.) International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR (2022), https://proceedings.mlr.press/v162/li22n.html 6
22. Li, M., Cai, W., Verspoor, K., Pan, S., Liang, X., Chang, X.: Cross-modal clinical graph transformer for ophthalmic report generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20656–20665 (2022) 4
23. 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, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 3334–3343. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00325, https://doi.org/10.1109/CVPR52729.2023.00325 2, 4, 6, 11
24. Li, M., Liu, R., Wang, F., Chang, X., Liang, X.: Auxiliary signal-guided knowledge encoder-decoder for medical report generation. World Wide Web pp. 1–18 (2022) 2
25. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out. Association for Computational Linguistics (Jul 2004) 10

26. Liu, F., Ge, S., Wu, X.: Competence-based multimodal curriculum learning for medical report generation. arXiv preprint arXiv:2206.14579 (2022) 11
27. 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) 2, 4, 11
28. Liu, F., Yin, C., Wu, X., Ge, S., Zhang, P., Sun, X.: Contrastive attention for automatic chest x-ray report generation. In: Findings of the Association for Computational Linguistics. pp. 269–280 (2021) 2, 6
29. Liu, X., Ji, K., Fu, Y., Du, Z., Yang, Z., Tang, J.: P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. CoRR **abs/2110.07602** (2021), https://arxiv.org/abs/2110.07602 5
30. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. pp. 3242–3250. IEEE Computer Society (2017). https://doi.org/10.1109/CVPR.2017.345, https://doi.org/10.1109/CVPR.2017.345 2
31. Mohsan, M.M., Akram, M.U., Rasool, G., Alghamdi, N.S., Abdullah-Al-Wadud, M., Abbas, M.: Vision transformer and language model based radiology report generation. IEEE Access **11**, 1814–1824 (2023). https://doi.org/10.1109/ACCESS.2022.3232719, https://doi.org/10.1109/ACCESS.2022.3232719 4
32. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (Jul 2002) 8, 10
33. Peng, Z., Hui, K.M., Liu, C., Zhou, B.: Learning to simulate self-driven particles system with coordinated policy optimization. Advances in Neural Information Processing Systems **34** (2021) 5
34. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019) 3, 4
35. Shin, T., Razeghi, Y., Logan IV, R.L., Wallace, E., Singh, S.: Autoprompt: Eliciting knowledge from language models with automatically generated prompts. arXiv preprint arXiv:2010.15980 (2020) 5
36. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 7433–7442. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.00718, https://doi.org/10.1109/CVPR52729.2023.00718 2, 4
37. Tanyel, T., Ayvaz, S., Keserci, B.: Beyond known reality: Exploiting counterfactual explanations for medical research. CoRR **abs/2307.02131** (2023). https://doi.org/10.48550/arXiv.2307.02131, https://doi.org/10.48550/arXiv.2307.02131 2
38. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 7
39. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P., Carroll, A., Lau, C., Tanno, R., Ktena, I., Mustafa, B., Chowdhery, A., Liu, Y., Kornblith, S., Fleet, D.J., Mansfield, P.A., Prakash, S., Wong, R., Virmani, S., Semturs, C., Mahdavi, S.S., Green, B., Dominowska, E., y Arcas, B.A., Barral, J.K., Webster, D.R., Corrado, G.S., Matias, Y., Singhal, K., Florence, P., Karthikesalingam, A., Natarajan, V.: Towards generalist biomedical AI. CoRR **abs/2307.14334** (2023).

https://doi.org/10.48550/arXiv.2307.14334, https://doi.org/10.48550/arXiv.2307.14334 9

40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 6
41. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015) 10
42. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015. pp. 3156–3164. IEEE Computer Society (2015). https://doi.org/10.1109/CVPR.2015.7298935, https://doi.org/10.1109/CVPR.2015.7298935 2
43. Virgolin, M., Fracaros, S.: On the robustness of sparse counterfactual explanations to adverse perturbations. Artif. Intell. **316**, 103840 (2023). https://doi.org/10.1016/j.artint.2022.103840, https://doi.org/10.1016/j.artint.2022.103840 2
44. Voutharoja, B.P., Wang, L., Zhou, L.: Automatic radiology report generation by learning with increasingly hard negatives. CoRR **abs/2305.07176** (2023). https://doi.org/10.48550/arXiv.2305.07176, https://doi.org/10.48550/arXiv.2305.07176 2
45. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023. pp. 11558–11567. IEEE (2023). https://doi.org/10.1109/CVPR52729.2023.01112, https://doi.org/10.1109/CVPR52729.2023.01112 2
46. Wang, Z., Zhou, L., Wang, L., Li, X.: A self-boosting framework for automated radiographic report generation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021. pp. 2433–2442. Computer Vision Foundation / IEEE (2021). https://doi.org/10.1109/CVPR46437.2021.00246, https://openaccess.thecvf.com/content/CVPR2021/html/Wang_A_Self-Boosting_Framework_for_Automated_Radiographic_Report_Generation_CVPR_2021_paper.html 2
47. Xu, D., Zhu, H., Huang, Y., Jin, Z., Ding, W., Li, H., Ran, M.: Vision-knowledge fusion model for multi-domain medical report generation. Inf. Fusion **97**, 101817 (2023). https://doi.org/10.1016/j.inffus.2023.101817, https://doi.org/10.1016/j.inffus.2023.101817 4
48. Yang, S., Wu, X., Ge, S., Zhou, S., Xiao, L.: Knowledge matters: Chest radiology report generation with general and specific knowledge. Medical Image Analysis **80**, 102510–102510 (2022) 4, 11
49. 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 Anal. **86**, 102798 (2023). https://doi.org/10.1016/j.media.2023.102798, https://doi.org/10.1016/j.media.2023.102798 2
50. 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 Anal. **86**, 102798 (2023). https://doi.org/10.1016/j.media.2023.102798, https://doi.org/10.1016/j.media.2023.102798 4

51. Yang, Z., Liu, Y., Ouyang, C., Ren, L., Wen, W.: Counterfactual can be strong in medical question and answering. Inf. Process. Manag. **60**(4), 103408 (2023). https://doi.org/10.1016/j.ipm.2023.103408, https://doi.org/10.1016/j.ipm.2023.103408 4
52. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12910–12917 (2020) 11
53. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16816–16825 (2022) 5
54. 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) 5