Visual-Linguistic Causal Intervention for Radiology Report Generation

Weixing Chen
Sun Yat-sen University
chen867820261@gmail.com

Yang Liu*
Sun Yat-sen University
liuy856@mail.sysu.edu.cn

Ce Wang
Institute of Computing Technology, Chinese Academy of Sciences
wangce@ict.ac.cn

Guanbin Li
Sun Yat-sen University
liguanbin@mail.sysu.edu.cn

Jiarui Zhu
Hong Kong Polytechnic University
zhujiarui42@gmail.com

Liang Lin
Sun Yat-sen University
linliang@ieee.org

Abstract

Automatic radiology report generation is essential for computer-aided diagnosis and medication guidance. Importantly, automatic radiology report generation (RRG) can relieve the heavy burden of radiologists by generating medical reports automatically from visual-linguistic data relations. However, due to the spurious correlations within image-text data induced by visual and linguistic biases, it is challenging to generate accurate reports that reliably describe abnormalities. Besides, the cross-modal confounder is usually unobservable and difficult to be eliminated explicitly. In this paper, we mitigate the cross-modal data bias for RRG from a new perspective, i.e., visual-linguistic causal intervention, and propose a novel Visual-Linguistic Causal Intervention (VLCI) framework for RRG, which consists of a visual deconfounding module (VDM) and a linguistic deconfounding module (LDM), to implicitly deconfound the visual-linguistic confounder by causal front-door intervention. Specifically, the VDM explores and disentangles the visual confounder from the patch-based local and global features without object detection due to the absence of universal clinic semantic extraction. Simultaneously, the LDM eliminates the linguistic confounder caused by salient visual features and high-frequency context without constructing specific dictionaries. Extensive experiments on IU-Xray and MIMIC-CXR datasets show that our VLCI outperforms the state-of-the-art RRG methods significantly. Source code and models are available at `https://github.com/WissingChen/VLCI`.

*Corresponding author is Yang Liu

1. Introduction

Radiology images (e.g., X-Ray, MRI) are widely used in clinical procedures, providing important evidence for disease analysis and medical intervention [47, 7]. Nevertheless, observing suspicious lesions and writing a coherent diagnosis report is time-consuming, even for experienced radiologists. Furthermore, inexperienced radiologists often fail to capture tiny abnormalities due to the high requirement for clinical knowledge. To relieve these issues [21], automatic radiology report generation (RRG) has emerged and attracted growing interest in recent years.

Similar to image captioning, RRG extracts the features from the medical images and generates a reliable report. However, the current RRG suffers from the following four challenges different from image captioning: 1) longer sentence generation, 2) more sophisticated linguistic and visual semantic patterns, 3) the abnormal regions within a radiology image are much smaller than that of the normal ones, and 4) the inherent anatomical structures within radiology images are challenging to be diagnosed, whereas entities in natural images are diverse and easily distinguishable. Therefore, these challenges present a substantial limit to modeling visual-linguistic interactions and learning informative cross-modal representations for accurate radiology report generation [3]. Great efforts have been devoted to solving these issues, such as additional knowledge [21], memory-driven module [3], and comparison with normal samples [22]. Actually, most of the previous methods aim to detect abnormalities, which can help discover potential lesions and list tiny abnormalities for accurate long-sequence generation. However, these methods usually focus on training computationally expensive models based on

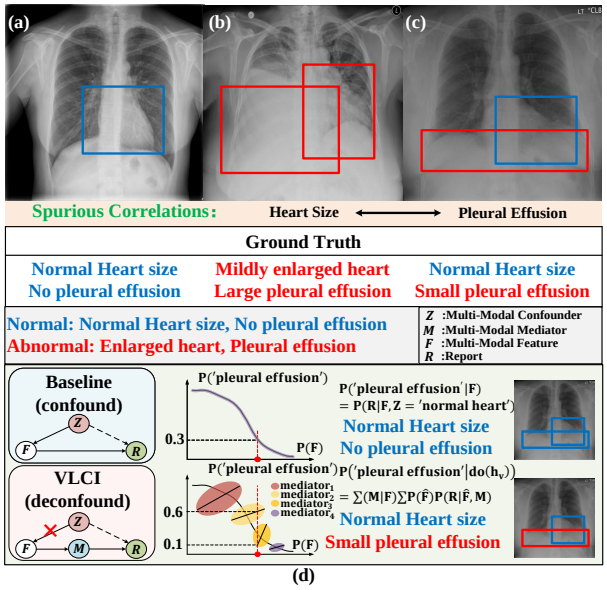


Figure 1. The visual-linguistic spurious correlation examples on IU-Xray dataset, where the colored texts in images are from the radiology reports and also describe as the colored box in radiology images. (a-b) shows the ground truth of two different RRG samples from the training set, and (c) is the sample from the testing set. The visual confounder (the visual feature of the heart) and linguistic confounder (the description of heart size) from (a-b) lead to spurious correlations and cause the wrong description of (c). (d) demonstrates the mechanism of causal intervention via the structural causal model (SCM). The Baseline tends to capture the spurious correlations and probability, while our VLCI can estimate the mediator by accumulating the probability of each sub-distribution and calculating the deconfounded probability of the correct word.

a large amount of samples and task-specific knowledge[1]. Actually, there exist significant visual and linguistic biases in numerous image-text data, as shown in Figure 1 (a-c). Therefore, lightweight models that can mitigate the cross-modal data bias are more essential for RRG to accurately discover abnormalities and generate reports [44, 27].

Actually, the most essential difficulty in abnormalities detection is the existence of the visual and linguistic biases that lead to entangled cross-modal features, i.e., spurious correlations, causing the incorrect report in the model prediction with confounders, as shown in Figure 1. Here, we can consider some high-frequently appearing concepts in the linguistic and visual modalities as the confounders. Specifically, the "enlarged heart" is frequently accompanied by "pleural effusion", leading to a spurious correlation between the multi-modal feature of the heart and the pleural, which causes the neglect of "small pleural effusion" in Figure 1. Different from the cascaded transformer-based feature extraction method [26], the mediator M (VLCI in Figure 1 (d)) can be considered as the intervention conditions that adjust the probability distribution of features to mitigate entanglement, rather than the direct feature extraction for report generation (Baseline in Figure 1 (d)). Therefore, it is challenging to generate accurate reports that reliably describe abnormalities due to the existence of confounder.

To mitigate visual-linguistic spurious correlations, causal intervention for image captioning often assume that the confounder is observable and alleviates the problem via back-door intervention, by approximating the observable confounder using a well-trained visual object detector or a well-constructed linguistic dictionary to cut off the shortcut path [36, 19]. Similarly, there also exists the data bias problem in RRG, and causal intervention can mitigate the visual-linguistic biases and improve the reliability of the generated report description. However, for RRG task, due to the complex data biases in visual and linguistic modalities, it is hard to represent the confounder explicitly. Fortunately, the front-door intervention gives a feasible way to calculate the confounder. Therefore, we utilize front-door intervention to implicitly mitigate cross-modal confounder and discover the true visual-linguistic causality by introducing an additional mediator involved in RRG [23, 43]. With front-door intervention, our model eliminates the spurious cross-modal correlations effectively and generates an accurate description of "small pleural effusion", as shown in Figure 1.

Motivated by the effectiveness of causal inference in deconfounding such cross-modal bias, we propose a lightweight cross-modal causal intervention framework for RRG without the observable confounder assumption, named Visual-Linguistic Causal Intervention (**VLCI**), to mitigate the visual and linguistic data biases. We combine Prefix Language Modeling (PLM) and Masked Image Modeling (MIM) for cross-modal feature alignment in pre-training. To mitigate the visual and linguistic biases, we propose the visual deconfounding module (**VDM**) and linguistic deconfounding module (**LDM**) based on the causal front-door intervention paradigm. The visual mediator is constructed by local detail information (e.g., lung texture) and global contour (e.g., pleural contour) from radiology images, targeting to discover and disentangle the visual feature. The linguistic confounder can be eliminated by the LDM, which estimates the change in the probability of word embedding caused by visual details and linguistic context. In summary, our main contributions are as follows:

- To implicitly mitigate cross-modal confounders and discover the true cross-modal causality, we propose visual-linguistic causal front-door intervention modules VDM and LDM. The VDM aims to disentangle the region-based features from images in the encoder, and the LDM aims to eliminate the spurious correlations caused by the visual-linguistic embedding.

[1]These methods build the template or knowledge database laboriously, making it hard to transfer those approaches directly to other datasets [41]

- To alleviate the problem of unpaired data when pre-training visual-linguistic RRG data, we combine the PLM and MIM for cross-modal pre-training in various data situations (e.g., unpaired, single modality), which is efficient and easy to implement.
- We propose a lightweight Visual-Linguistic Causal Intervention (VLCI) framework for RRG, which introduces mediators without additional knowledge, to implicitly deconfound the visual-linguistic confounder by causal front-door intervention. Experimental results show that VLCI achieves state-of-the-art performance on two datasets IU-Xray and MIMIC-CXR.

2. Related Work

2.1. Image Captioning

Image captioning aims to understand image information and describe it in text, which mainly adopts the encoder-decoder framework [33]. Generally, image features extracted by the encoder are fed into the decoder, often based on recurrent neural networks (RNN) and transformers [6]. The recent work achieved great success in this task [1, 10, 14, 18], which presented the spatial relationships of regional features, and rely on the integration of visual and semantic data to improve performance. Compared with the image captioning approaches, the RRG has similar structures [33]. Nevertheless, image captioning usually generates a single sentence to describe the main entities, while the RRG focuses on the potential subtle abnormalities areas in medical images and generate longer sentence from more sophisticated visual-linguistic semantics.

2.2. Radiology Report Generation

Recently, RRG methods have followed the works of image captioning and have shown remarkable performance. For the issues above, the knowledge-aware module [21, 41, 38], template retrieval module [22], and memory-driven module [3, 27] are used to generate useful reports. However, it still has some limitations. For the data bias, PPKED [21] and RG-GSK [41] explored and distilled the different kinds of knowledge for RRG, which needs to be annotated. Following that, CA [22] and CMCL [20] utilized the comparison of data differences to enact the training strategy, which needs more data to estimate distribution. Chen et al. proposed to generate reports with a memory-driven transformer but inference with slow speed [3]. Moreover, M2TR [27] and MSAT [38] integrate memory into attention, while needing more computational resources to generate reports. In summary, the visual-linguistic bias hinders the promotion of the application of RRG, while our lightweight VLCI implicitly mitigates cross-modal confounders and discovers the true cross-modal causality by causal front-door intervention and reducing the dependency on additional annotation for discovering the abnormalities.

2.3. Causal Inference

Causality provides a new methodology to design robust models via the elimination of spurious correlation [9, 29, 24]. Causal inference estimates the hidden causal effects in the distribution while significantly improving the model's generalization. It mitigates confounders through back-door, front-door intervention, or counterfactual intervention [42]. For example, Wang et al. [36] improved Faster R-CNN by causal back-door intervention to obtain a more robust object detection model, which improves the performance of VQA and image captioning. However, the confounder is usually unobservable and elusive, thus front-door intervention and counterfactual intervention can be applied [43, 37]. Therefore, causal inference has achieved remarkable performance in cross-modal tasks [23, 19]. Compared with the previous works that address VQA or image captioning, we focus on radiology report generation and propose visual-linguistic causal intervention, which simultaneously eliminates spurious correlations from visual and linguistic confounders.

3. Method

In this section, we first introduce the pre-training strategy, followed by the two essential cross-modal causal intervention modules, i.e., the Visual Deconfounding Module (VDM) and the Linguistic Deconfounding Module (LDM). Next, we describe how to integrate these two modules into the VLCI for cross-modal causal intervention.

3.1. Overview

A typical RRG model takes a radiology image $I \in \mathbb{R}^{C\times H\times W}$ as input and generates the corresponding report $R = \{w_1, w_2, \ldots, w_n\}$ that contains critical information. As illustrated in Figure 2, the VLCI employs the transformer structure to model $P(R|I) = \sum_{i=1}^{n} P(w_i|h_v, h_w)$, where h_v is the visual feature extracted by an encoder and guides the prefix word embedding h_w to generate the next word w_i with visual-linguistic deconfounding. To ensure that the estimation of confounder Z is caused by the prior $P(I)$ and $P(R)$, we leverage the Visual-Language Pre-training (VLP) model to construct the correlation between the visual contexts and linguistic concepts. Meanwhile, due to the absence of a knowledge graph and a well-trained feature extractor. We innovatively leverage causal front-door intervention to eliminate implicitly the spurious correlations from visual and linguistic modalities, and it is integrated into VDM and LDM, respectively.

3.2. Visual-Linguistic Pre-training (VLP)

In the medical pre-training framework, there exist two difficulties: (1) The unpaired data that only has a single modality is hard to be utilized in supervised learning, (2) heterogeneous data that makes it difficult to distinguish

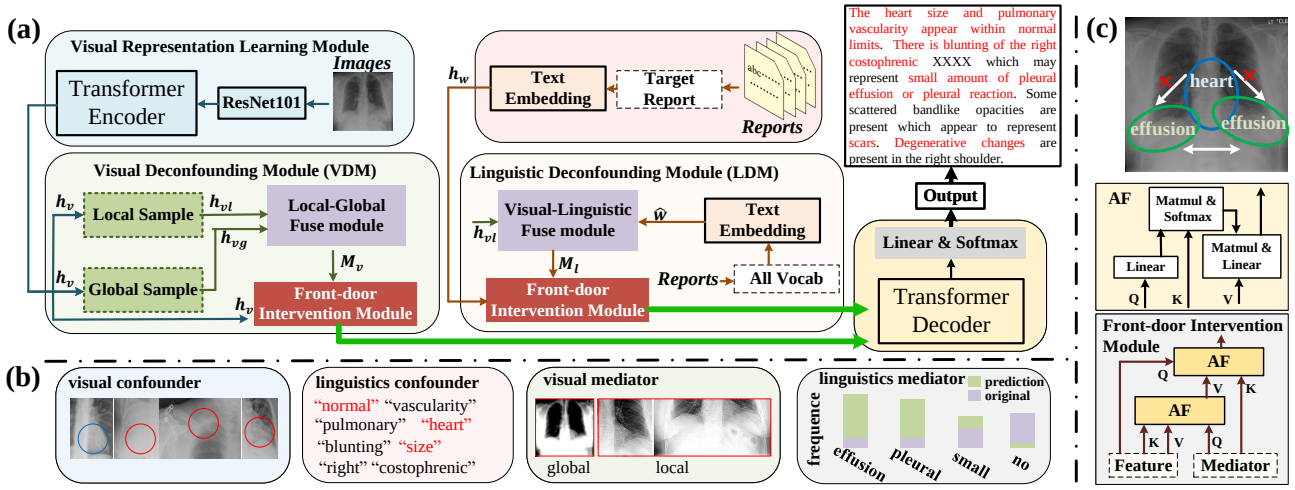


Figure 2. The overview of our Visual-Linguistic Causal Intervention (VLCI) framework, which consists of a Visual Representation Learning Module (VRLM), a Visual Deconfounding Module (VDM), and a Linguistic Deconfounding Module (LDM). The ResNet backbone uses the first three blocks of ResNet101, and the text embedding layers are weight-sharing. Specifically, the VDM explores visual bias via local sampling and global sampling. The LDM estimates linguistic bias via a vocab dictionary and visual features.

the region feature because the morphology of the same lesion varies greatly [47]. Since the cross-modal pre-training provides fine-grained regional features without regional label [46], we utilize PLM and MIM in linguistic and visual modeling to deal with unpaired data. Therefore, we use a single encoder to extract multi-modal features and two weight-shared decoders to solve PLM and MIM tasks, respectively [39, 12]. In each block of the multi-modal encoder, the attention layer is weight-shared while the two feed-forward layers handle the corresponding modal feature respectively [45] (Refer to Appendix A).

Motivated by SimVLM [39], we extract image features from the first three blocks of ResNet101 [13] as prefix tokens in PLM. Simultaneously, the text is divided randomly into two parts, of which one is generated by another under the guidance of the obtained image tokens. When the corresponding image is absent, the PLM can also be trained with only text modality, which is the same as SimVLM. Assume that $h_v \in \mathbb{R}^{\frac{HW}{P^2} \times d}$ is denoted as the image token extracted by the raw image I, where P is the patch size, and d is the embedding size. Then$\{w_{n_p}, \ldots, w_n\}$ is the postfix sequence after the textual description h_w of length $n_p \geq 0$. Thus, the formulation is as follows:

$$\mathcal{L}_{\text{PLM}}(\theta) = -\sum_{i=n_p}^{n} log P_\theta(w_i | h_v, h_{w_{<n_p}}), \qquad (1)$$

where θ is the trainable parameters of the model, h_v is the visual embedding with a trainable 2D positional encoding, h_w is learned for a fixed vocabulary and received by the encoder as the prefix, and n is the report length.

To deal with unpaired images like MAE [12], we take

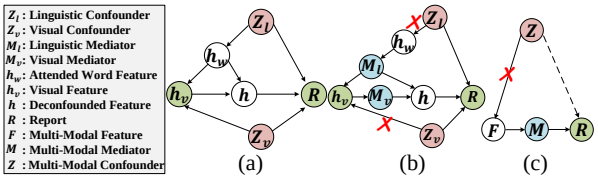


Figure 3. Front-door causal intervention $P(R|do(h_v), do(h_w))$ is implemented by the mediator M_v and M_l. It is the SCM of RRG with the confounder Z_v and Z_l in (a) and cut off the path $Z_v \rightarrow h_v$ and $Z_l \rightarrow h_w$ via blocking the back-door path $h_v \leftarrow Z_v \rightarrow R$ and $h_v \leftarrow h_w \leftarrow Z_l \rightarrow R$ in (b) and get the SCM of VLCI in (c).

advantage of the MIM paradigm. Additionally, since the MIM is trained with pairwise data, the missing semantics of masked images can be provided by text to enhance the cross-modal association [8]. Thus, we reconstruct the masked visual token via the semantics of the unmasking visual token and linguistic token, which can learn the tiny difference in the dataset [32]. The target of the MIM can be formulated as follows:

$$\mathcal{L}_{\text{MIM}}(\theta) = P_\theta(h_{vm} | h_{vv}, h_w), \qquad (2)$$

where h_{vm} denotes the masked visual tokens extracted by the ResNet backbone, h_{vv} is the unmasked tokens, and h_w denotes the word tokens of the whole report. Then the ResNet and the multi-modal transformer encoder are utilized as VRLM in the downstream tasks.

3.3. Visual-Linguistic Causal Intervention

After visual-linguistic pre-training, the trained visual-linguistic feature encoders still contain visual and linguistic biases from cross-modal confounders [42]. Therefore,

we employ Pearl's structural causal model (SCM) [9] to characterize the causal effect between visual and linguistic modalities when conducting RRG. As illustrated in Figure 3 (a), the causal effects $h_v \rightarrow R$ and $h_w \rightarrow R$ are affected by the confounder $Z = \{Z_v, Z_l\}$ from back-door paths $h_v \leftarrow Z_v \rightarrow R$ and $h_w \leftarrow Z_l \rightarrow R$ [19], respectively. In our SCM, the non-interventional prediction can be expressed by the Bayes rule:

$$\begin{aligned} P(R|I) &= P(R|h_v, h_w) \\ &= \sum_{i=1}^{n} \sum_{z} P(w_i|h_v, h_w, Z=z) P(Z=z|h_v, h_w), \end{aligned} \tag{3}$$

where Z brings the spurious correlation via $P(Z = z|h_v, h_w)$, leading to incorrect reports. Taking the example in Figure 1, when $P(Z =$ "normal heart"$|h_v =$ "heart", $h_w =$ "normal") is large while $P(Z =$ "enlarged heart"$|h_v =$ "heart", $h_w =$ "normal") is small, it will enlarge $P(R =$ "no pleural effusion"$|h_v, h_w, Z =$ "normal heart"). To mitigate visual-linguistic confounders and uncover the cross-modal causal structure, we apply causal front-door intervention by introducing mediator M_v and M_l, respectively, as shown in Figure 3(b). Generally, Z_v is unobservable without a well-trained object detector, and the back-door path $h_v \leftarrow Z_v \rightarrow R$ can be blocked by M_v via learning the true causal effect $h \leftarrow M_v \leftarrow h_v \leftarrow Z_v \rightarrow R$. The confounders and mediator can be shown intuitively in Figure2 (b). Similarly, the intervention on the back-door path $h_v \leftarrow h_w \leftarrow Z_l \rightarrow R$ can be implemented by calculating the M_l without well-constructed confounder dictionaries. Since the front-door intervention can eliminate unobservable confounders, we integrate the Front-door Intervention Module (FIM) into VDM and LDM.

3.3.1 Front-door Intervention Module (FIM)

To cut off back-door paths $h_v \leftarrow Z_v \rightarrow R$ and $h_w \leftarrow Z_l \rightarrow R$ via M_v and M_l (SCM in Figure 3(b)), we leverage the do calculus $do(\cdot)$ [19, 23, 25, 30], which is formulated as:

$$\begin{aligned} &P(R|do(h_v), do(h_w)) = \\ &\sum_{m} P(R|do(M=m)) P(M=m|do(h_v), do(h_w)), \end{aligned} \tag{4}$$

where M is the mediator containing M_v and M_l. Since the intervention probability is equal to the conditional probability in the path $\{h_v, h_w\} \rightarrow M$, the back-door path between M and R can be blocked, and enlarge $P(R =$ "small pleural effusion"$|h_v =$ "heart", $h_w =$ "normal", $M =$ "pleural"), as the previously mentioned example, where the "pleural" is the multi-modal feature of pleural in each condition of heart size. Consequently, we can assume that $\{h_v, h_w\}$ is feature F, and formulate Eq. (4) as:

$$\begin{aligned} &P(R|do(h_v), do(h_w)) = P(R|do(F)) = \\ &\sum_{m} P(M=m|F) \sum_{\hat{F}} P(F=\hat{F}) P(R|F=\hat{F}, M=m). \end{aligned} \tag{5}$$

To further estimate $P(R|do(F))$, we implement the front-door causal intervention Eq. (5) with the deep learning framework. Here, we adopt Normalized Weighted Geometric Mean (NWGM) [40] and approximate the Eq. (5) as:

$$P(R|do(h_v), do(h_w)) \approx \text{Softmax}(g(h_w, h_v, \hat{M}_v, \hat{M}_l)), \tag{6}$$

where $g(\cdot)$ denote the network mapping functions, $\hat{M}_v$ and $\hat{M}_l$ denote the estimations of M_v and M_l via VDM and LDM. In Figure 2 (c), the FIM consists of two Attention Fusion (AF) layers and it is integrated into VDM and LDM.

3.3.2 Visual Deconfounding Module (VDM)

In Figure 2, we calculate the visual mediator M_v via local feature h_{vl} and global feature h_{vg}. The h_{vl} is denoted as the local detail information acquired from Local Sampling, while the h_{vg} is the contours and position feature acquired from Global Sampling [34]. For instance, the contour of the heart affects the determination of pleural effusion, and the texture of the lungs can also be the basis of detection.

Local Sampling. Inspire by TransFG [11], we use the attention accumulated from the encoder to select top k tokens that correspond to the report and only use these selected tokens as $h_{vl} \in \mathbb{R}^{k \times d}$, where $k = 6$ for each head of attention. Then, h_{vl} is enhanced via CaaM [37], which further excavates the local internal relations. The h_{vl} aims to obtain local critical details in the image, which can be used as the key basis for RRG.

Global Sampling. The global sampling is implemented by Down Sampling Transformer block, in which the 14×14 visual tokens are down-sampled to 7×7 as $h_{vg} \in \mathbb{R}^{49 \times d}$. Max pooling in this block can better retain the global structure information in the image as the general features of the data itself. We formulate the operation as follows:

$$h_{vg} = W[\text{P}(h_v) + \text{Attn}(\text{P}(\text{LN}(h_v))], \tag{7}$$

where P is the 2d max pooling layer, LN is layer normalization, Attn is the 2d relative attention [4], and W denotes the weights of linear layer.

Finally, the h_{vl} is integrated with h_{vg} to enhance local details with global structural information via Local-Global Fuse Module formulated as Eq. (8), namely mediator M_v.

$$M_v = \text{FFN}([\text{MHA}(h_{vl}, h_{vl}, h_{vl}), \text{MHA}(h_{vl}, h_{vg}, h_{vg})]) \tag{8}$$

where MHA and FFN are the Multi-Head Attention layer and Feed-Forward Network layer, respectively. $[\cdot, \cdot]$ denotes concatenation.

	Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	Rouge-L	METEOR	Precision	Recall	F1
						IU-Xray Dataset					
Lightweight	R2Gen[3]	0.470	0.304	0.219	0.165	/	0.371	0.187	/	/	/
	CMCL[20]	0.473	0.305	0.217	0.162	/	0.378	0.186	/	/	/
	PPKED[21]	0.483	0.315	0.224	0.168	0.351	0.376	0.190	/	/	/
	CA[22]	0.492	0.314	0.222	0.169	/	0.381	0.193	/	/	/
	AlignTransformer[44]	0.484	0.313	0.225	0.173	/	0.379	0.204	/	/	/
Heavyweight	M2TR[27]	0.486	0.317	0.232	0.173	/	**0.390**	0.192	/	/	/
	RG-GSK[41]	**0.496**	**0.327**	0.238	0.178	0.382	0.381	/	/	/	/
	VLCI (ours)	0.495	**0.327**	**0.239**	**0.185**	**0.449**	0.389	**0.206**	/	/	/
						MIMIC-CXR Dataset					
Lightweight	R2Gen[3]	0.353	0.218	0.145	0.103	/	0.277	0.142	0.333	0.273	0.276
	CMCL[20]	0.334	0.217	0.140	0.097	/	0.281	0.133	/	/	/
	PPKED[21]	0.360	0.224	0.149	0.106	0.237	0.284	0.149	/	/	/
	CA[22]	0.350	0.219	0.152	0.109	/	0.283	0.151	0.352	0.298	0.303
	AlignTransformer[44]	0.378	0.235	0.156	0.112	/	0.283	0.158	/	/	/
Heavyweight	M2TR[27]	0.378	0.232	0.154	0.107	/	0.272	0.145	0.240	**0.428**	0.308
	RG-GSK[41]	0.363	0.228	0.156	0.115	0.203	0.284	/	**0.458**	0.348	0.371
	MSAT*[38]	0.373	0.235	0.162	**0.120**	**0.299**	0.298	0.143	/	/	/
	VLCI (Ours)	**0.390**	**0.248**	**0.167**	0.119	0.168	**0.302**	**0.172**	0.409	0.390	**0.398**

Table 1. The performances of VLCI and other methods on IU-Xray and MIMIC-CXR datasets. The 1^{st} and 2^{nd} best results are bolded and underlined, respectively. The method marked by * means its result is from [38], while the rest is from [41], and / means the absent result.

3.3.3 Linguistic Deconfounding Module (LDM)

For linguistic deconfounding, we have some observations from the link $h_v \leftarrow h_t \leftarrow Z_l \rightarrow R$: (1) the linguistic contexts can affect the generation of the next word, and (2) the attended word features affect the attended visual features via cross-attention [19]. Additionally, the difference in word frequency brings a large distance deviation in embedding space, so the distance of word vectors cannot represent semantic relevance well [17]. Thus, we calculate the linguistic mediator M_l in embedding space via all vocabularies from the tokenizer as the global feature and use h_{vl} obtained from the VDM, which estimates the current word frequency to adjust the distribution of h_w (Figure 2).

$$\begin{aligned} h_{vl}^{'} &= \text{FFN}(\text{MHA}(h_{vl}, \hat{w}, \hat{w})); \\ M_t &= \text{FFN}(\text{MHA}(h_{vl}^{'}, h_{vl}, h_{vl})) \end{aligned} \tag{9}$$

where $\hat{w}$ denotes all word tokens from the tokenizer. Then, we build the causal link $h \leftarrow M_l \leftarrow ht \leftarrow Z_l \rightarrow R$ to cut off the path $Z_l \rightarrow h_v$ via M_l.

In Figure 2, the deconfounded visual and linguistic features are fed to the decoder to learn fused cross-modal features. The output layer is a linear projection with softmax operation, mapping probability into N-dimensional, where N is the vocabulary size. Finally, the training target is to minimize the negative log-likelihood loss:

$$\mathcal{L}_{\text{nll}}(\theta) = -\sum_{i=1}^{n} log[P_\theta(w_i|do(h_v), do(h_{w_{<i}}))] \tag{10}$$

4. Experiment

4.1. Experimental settings

Dataset. **IU-Xray** [5], namely Indiana University Chest X-ray Collection, is a public radiology dataset widely used to evaluate the performance of RRG methods. It contains 7,470 chest images and 3,955 corresponding reports. We apply the same setting as R2Gen [3] and tokenize the words more than three occurrences. **MIMIC-CXR** [16] is a large-scale chest radiology dataset, with 377,110 images and 227,835 corresponding reports. We use the official split and tokenize the words with more than ten occurrences.

Evaluation Metrics. We adopt the widely used NLG metrics, including BLEU [28], ROUGE-L [31], METEOR [2] and CIDEr [35]. Since the RRG specifically focuses on the abnormality detection accuracy rather than the text fluency and similarity with the real report, we further adopt clinical efficacy (CE) metrics [3, 22, 27, 41]. It is calculated by the labels extracted from CheXpert [15].

Implementation Settings. We use the first three blocks of ResNet101 [13] to extract 1,024 feature maps, which are projected into 512 maps of size 14×14. The dimension of the latter VLCI layers and the number of attention heads are fixed to 512 and 8. The number of layers is 3 for both the encoder and the decoder. We adopt the same dataset during pretraining and fine-tuning. The batch size is set to 64 in pretraining and 16 in fine-tuning. In the pre-training stage, we adopt the image mask rate of 85% (Refer to Appendix B) for MIM. The VLP is trained by the AdamW optimizer with a warm-up step 10% of the whole train step and the peak learning rate is 5e-4. The weight decay of the optimizer is set to 1e-2. The total epochs are set to 100 and 30 for the IU-Xray and MIMIC-CXR datasets, respectively. The model is

Type	Abnormalities	Baseline	VLCI
Lung	Lung Opacity	0.497	0.542 (0.045)
	Lung Lesion	0.924	0.929 (0.005)
	Consolidation	0.617	0.862 (0.245)
	Pneumonia	0.639	0.805 (0.166)
	Atelectasis	0.657	0.673 (0.016)
	Edema	0.509	0.746 (0.237)
Pleural	Pneumothorax	0.564	0.940 (0.376)
	Pleural Effusion	0.551	0.727 (0.176)
	Pleural Other	0.957	0.958 (0.001)
Other	Enlarged Cardiomediastinum	0.473	0.632 (0.159)
	Cardiomegaly	0.454	0.598 (0.144)
	Fracture	0.924	0.933 (0.009)
	Support Devices	0.761	0.701 (-0.060)

Table 2. Evaluation of abnormality classification results (accuracy) on MIMIC-CXR. The numbers in the bracket mean the improvement with green and decrement with red.

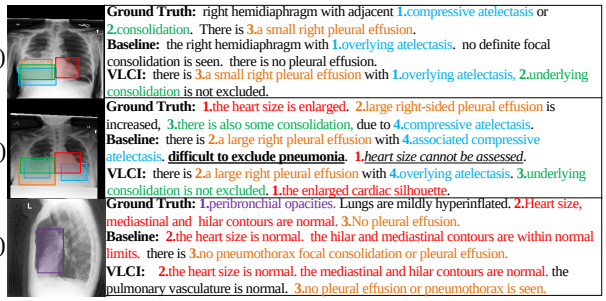


Figure 4. Examples of generated reports on MIMIC-CXR. The baseline is a transformer with the same setting as our VLCI. Different colors are applied to the target keywords. The uncertain and wrong words are underlined with italics and bold, respectively.

finetuned by the Adam optimizer with an initial learning rate of 1e-5 and weight decay of 5e-5 for 10 and 3 epochs for the IU-Xray and MIMIC-CXR dataset, respectively.

Baseline Models. We compare the proposed VLCI model with several state-of-the-art RRG models, which are divided into lightweight and heavyweight models (Refer to Appendix C). Specifically, the lightweight models have no more than 3-layer modules for both the encoder and the decoder. Most of them employ different modules for boosting model performances, which are computationally expensive. Note that the total parameters of our VLCI are comparable to the R2Gen, while VLCI is faster because our model gets rid of the dependency of calculating memory recursively.

4.2. Quantitative Analysis

As shown in Table 1, our VLCI outperforms almost all the RRG methods. Specifically, compared with the lightweight AlignTransformer, the VLCI significantly boosts the BLEU-4 metric by 1.2% on the IU-Xray dataset and 0.7% on the MIMIC-CXR dataset. Compared with heavyweight model M2TR, our lightweight VLCI boosts the METEOR metric by 1.4% and 2.7% on the IU-Xray and MIMIC-CXR datasets, respectively. Since the METEOR metric considers the synonyms, it shows the effectiveness of our causal intervention for discovering semantic correlation. But the performance of VLCI is slightly lower than RG-GSK on the IU-Xray dataset for the BLEU-1 metric, due to the fact that BLEU-1 metric evaluates the single word performance of the reports. While our performance for the BLEU-1 metric is significantly improved on MIMIC-CXR dataset. It reveals that the visual-linguistic data bias and the spurious correlation among multiple words are more significant on large dataset MIMIC-CXR, which can fully take advantage of our cross-modal causal intervention.

Similarly, for the Rough-L metric that calculates the recall of each word, the performance of VLCI and M2TR shows the similar phenomena. However, for the CIDEr and BLEU-4 metrics, the performance of VLCI is lower than that of MSAT on MIMIC-CXR. These metrics tend to evaluate the similarity of the whole report, which is challenging for a long-sequence report with medical terminology. Thus, the knowledge-based approach (PPKED, RG-GSK, MSAT) with professional concepts can generate a more precise description of the disease and achieve better performance, while VLCI only intervenes in the causal effect without external knowledge. The CE metric is only applied to MIMIC-CXR dataset because the label extractor (CheXpert) [15] is specially designed for MIMIC-CXR to obtain class labels. Compared with the state-of-the-art lightweight CA in Table 1, VLCI improves the performance by 5.7% in Precision, 9.2% in Recall, and 9.5% in F1-Score. This validates that VLCI can provide a more accurate clinic diagnosis rather than only generating a fluent report.

4.3. Qualitative Analysis

To further analyze the clinic diagnosis from VLCI, we evaluate the abnormality used in the CE metric. In Table 2, VLCI boosts the performance of the abnormalities detection on the MIMIC-CXR dataset, especially the accuracy of "Pneumothorax", "Edema" and "Consolidation". This is because our VLCI explores sufficient visual information and further produces more accurate and less biased descriptions by cross-modal causal intervention than the Transformer baseline. However, the estimation of some categories still keeps ambiguous, e.g., "Lung Opacity". It reveals that VLCI can provide an comprehensive consideration of various radiologic signs to detect the abnormality but give less improvement for the single source abnormality. For example, whether "Edema" is caused by the heart has different radiologic signs, while the increase in lung density can be considered as "Lung Opacity". Thus, VLCI can capture the abnormality with complex causes more effectively, where exists more spurious correlations. Besides, Table 2 shows the unavailability of causal intervention in independent abnormalities, e.g. "Support Devices".

We further conduct the qualitative analysis on MIMIC-CXR dataset via three intuitive generated examples of the

Method	BLEU-1	BLEU-4	CIDEr	Rouge-L
Baseline	0.433	0.148	0.501	0.345
w/ MAE	0.449	0.154	0.486	0.360
w/ VLP (MIM)	0.439	0.162	**0.602**	0.362
w/ VLP (PLM)	**0.467**	**0.165**	0.538	**0.365**
w/ VLP (PLM+MIM)	0.466	0.160	0.431	0.364
w/ VLP (PLM)*	0.448	0.151	0.399	0.349
w/ VLP (PLM+MIM)* (Ours)	0.452	0.161	0.522	0.351

Table 3. The performance of different pre-training methods on IU-Xray, the result marker by * means finetuning on downstream task with 5 epoch, while the rest only use the encoder with 100 epochs.

baseline and the VLCI in Figure 4. Particularly, as in Figure 4 (a), the reference report consists of three abnormalities, the baseline neglects "pleural effusion" and "consolidation", while VLCI discovers all abnormalities accurately. It shows that our VDM can comprehensively perceive all essential visual features. Figure 4 (b) shows an example where the same visual region is simultaneously discovered by the baseline and the VLCI, but leads to different descriptions. Our VLCI can accurately describe the heart, while the baseline is uncertain and even has a miscalculation of pneumonia. It shows that LDM can alleviate the semantic bias caused by word frequency in word embedding space. Figure 4 (c) shows a normal case that only contains "Lung Opacity". Both the Baseline and VLCI can generate a fluent report and indicates the normality. But VLCI fails to capture the peribronchial opacities, which are the radiologic signs between "Clear Lung" and "Consolidation". This is because the "Lung Opacity" only changes the pulmonary density and it is difficult to be discovered determinedly.

4.4. Ablation Studies

4.4.1 Effectiveness of VLP

In Table 3, we make a comparison with different pre-training methods. It shows that the cross-modal pre-training method has a more robust representation ability than the MIM with single-modality. Additionally, our cross-modal pre-training achieves comparable performance to the PLM model that only finetunes the encoder, while ours finetunes the whole model with fewer epochs.

Furthermore, in Table 4, Baseline$^{w\blacklozenge\bullet}$ is significantly worse than baseline on MIMIC-CXR dataset, e.g., 0.101 $\rightarrow$ 0.070 for BLEU-4, while still keeping performance improvement on IU-Xray dataset. This validates the significant feature complexity from the large-scale MIMIC-CXR dataset leads to unstable probability distribution estimation with causal intervention. Meanwhile, we find that the VLP can substantially boost the performance of the baseline, e.g., 0.148 $\rightarrow$ 0.161, 0.101 $\rightarrow$ 0.108 for BLEU-4 on IU-Xray and MIMIC-CXR datasets, respectively. The improvement is caused by the learned comprehensive concepts and context in the pre-training and the cross-modal feature alignment stage, which shows the importance of VLP. Similarly,

Dataset	Method	BLEU-4	CIDEr	Rouge-L
IU-Xray	Baseline	0.148	0.501	0.345
	Baseline$^{w\blacklozenge}$ (w/ VDM)	0.160	0.521	0.364
	Baseline$^{w\bullet}$ (w/ LDM)	0.155	0.509	0.361
	Baseline$^{w\blacklozenge\bullet}$ (w/ VDM&LDM)	0.163	0.544	0.361
	R2Gen	0.165	0.493	0.360
	R2Gen$^{w\blacklozenge}$ (w/ VDM)	0.171	0.553	0.370
	R2Gen$^{w\bullet}$ (w/ LDM)	0.166	0.546	0.360
	R2Gen$^{w\blacklozenge\bullet}$ (w/ VDM&LDM)	0.173	**0.628**	0.368
	Baseline$^{w\bigstar}$ (w/ VLP)	0.161	0.522	0.351
	Baseline$^{w\bigstar\blacklozenge}$ (w/ VLP&VDM)	0.176	0.514	0.377
	Baseline$^{w\bigstar\bullet}$ (w/ VLP&LDM)	0.175	0.342	0.382
	VLCI	**0.185**	0.449	**0.389**
MIMIC-CXR	Baseline	0.101	0.137	0.274
	Baseline$^{w\blacklozenge\bullet}$ (w/ VDM&LDM)	0.070	0.074	0.230
	Baseline$^{w\bigstar}$ (w/ VLP)	0.108	0.167	0.298
	Baseline$^{w\bigstar\blacklozenge}$ (w/ VLP&VDM)	0.110	0.157	0.297
	Baseline$^{w\bigstar\bullet}$ (w/ VLP&LDM)	0.117	0.158	0.299
	VLCI	**0.119**	**0.168**	**0.302**

Table 4. Ablation analysis of our VLCI. The Baseline is implemented by the transformer. The marker at Baseline and R2Gen [3] means the operation in the brackets.

The Rough-L is also barely improved due to the feature complexity and long sequence from MIMIC-CXR dataset. For example, although AlignTransformer achieves the same score of the Rough-L as CA on MIMIC-CXR dataset, it outperforms CA on all other metrics.

4.4.2 Effectiveness of Causal Intervention

VDM. In Table 4, Baseline$^{w\blacklozenge}$ and R2Gen$^{w\blacklozenge}$ can boost the performance compared to Baseline and R2Gen, which demonstrates the model-agnostic property of the VDM. However, the improvement of BLEU-4 between Baseline$^{w\bigstar\blacklozenge}$ and Baseline$^{w\bigstar}$ on IU-Xray dataset is more significant than that on MIMIC-CXR dataset. This is because the VDM can discover more essential visual information, but the report of the MIMIC-CXR dataset is more complex and the model fails to generate accurate descriptions. The performance degradation of CIDEr can further illustrate it. In Figure 5, the attention map from the encoder of our VLCI can truly focus on the dominated area of possible abnormalities rather than spurious correlations with biased visual concepts. This validates that the VDM is semantics-sensitive to capture dominant visual content by conducting visual causal intervention.

LDM. Compared to the VDM, the LDM plays a more significant role in RRG because the sophisticated linguistic semantic patterns within reports are entangled and biased that require elaborate linguistic deconfounding. In Table 4, the performance drops without LDM, e.g., 0.119 $\rightarrow$ 0.110 for the BLEU-4 metric on MIMIC-CXR dataset. This shows the importance of adjusting semantic relevance in word embedding space. Compared with the baseline, the performance improvement of Baseline$^{w\bigstar\bullet}$ on MIMIC-CXR dataset demonstrates that the LDM can generate more ac-

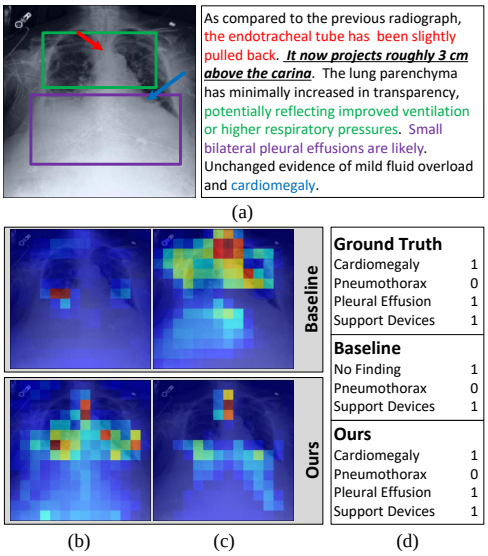


Figure 5. The visualization of the attention map. (a) is an example from MIMIC-CXR dataset that the colored text should be discovered in the marked region of the image. The images in (b-c) are the attention maps of the baseline and our VLCI, respectively. (b) are the accumulated attention maps from the encoder for the selected local feature, and (c) are the response to the "pleural" (decoder output). The labels shown in (d) are extracted by CheXpert.

curate reports even with biased visual information. However, the CIDEr metric on IU-Xray dataset shows the effectiveness of the combination of VDM and LDM, while ILVD obtains a lower score. This is due to the worse diversity on IU-Xray dataset, where Baseline$^{w\blacklozenge\bullet}$ and R2Gen$^{w\blacklozenge\bullet}$ can get higher CIDEr but lower BLEU-4 with inadequate multi-modal feature correlation. In Figure 5 (c), the attention map of the baseline decoder shows an obvious redundancy response, while VLCI can capture dominated semantic information in a coarse-to-fine manner, which is more related to the abnormalities. These results show that LDM can capture more discriminative semantic information from linguistic modality by linguistic front-door intervention.

5. Conclusion

In this paper, we propose Visual-Linguistic Causal Intervention (VLCI) framework for RRG, to implicitly deconfound the visual-linguistic confounder by causal front-door intervention. To alleviate the problem of unpaired visual-linguistic data when pre-training, we combine the PLM and MIM for cross-modal pre-training. To implicitly mitigate cross-modal confounders and discover the true cross-modal causality, we propose visual-linguistic causal front-door intervention modules VDM and LDM. Experiments on IU-Xray and MIMIC-CXR datasets show that our VLCI can effectively mitigate visual-linguistic bias and outperforms the state-of-the-art methods. The lower computational cost and faster inference speed of VLCI promote its clinical application. We believe our work could inspire more causal reasoning methods in medical report generation.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018.

[2] Satanjeev Banerjee and Alon Lavie. 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*, pages 65–72, 2005.

[3] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020.

[4] Zihang Dai, Hanxiao Liu, Quoc V Le, and Mingxing Tan. Coatnet: Marrying convolution and attention for all data sizes. *NeurIPS*, 34:3965–3977, 2021.

[5] 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, 2015.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] Joann G Elmore, Gary M Longton, Patricia A Carney, Berta M Geller, Tracy Onega, Anna NA Tosteson, Heidi D Nelson, Margaret S Pepe, Kimberly H Allison, Stuart J Schnitt, et al. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Jama*, 313(11):1122–1132, 2015.

[8] Xinyang Geng, Hao Liu, Lisa Lee, Dale Schuurams, Sergey Levine, and Pieter Abbeel. Multimodal masked autoencoders learn transferable representations. *arXiv preprint arXiv:2205.14204*, 2022.

[9] Madelyn Glymour, Judea Pearl, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[10] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. Normalized and geometry-aware self-attention network for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10327–10336, 2020.

[11] 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 *AAAI*, volume 36, pages 852–860, 2022.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[14] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. *Advances in Neural Information Processing Systems*, 32, 2019.

[15] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, volume 33, pages 590–597, 2019.

[16] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[17] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. *arXiv preprint arXiv:2011.05864*, 2020.

[18] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8928–8937, 2019.

[19] Bing Liu, Dong Wang, Xu Yang, Yong Zhou, Rui Yao, Zhiwen Shao, and Jiaqi Zhao. Show, deconfound and tell: Image captioning with causal inference. In *CVPR*, pages 18041–18050, 2022.

[20] Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. *arXiv preprint arXiv:2206.14579*, 2022.

[21] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *CVPR*, pages 13753–13762, 2021.

[22] Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. Contrastive attention for automatic chest x-ray report generation. *arXiv preprint arXiv:2106.06965*, 2021.

[23] Yang Liu, Guanbin Li, and Liang Lin. Cross-modal causal relational reasoning for event-level visual question answering. *arXiv preprint arXiv:2207.12647*, 2022.

[24] Yang Liu, Yu-Shen Wei, Hong Yan, Guan-Bin Li, and Liang Lin. Causal reasoning meets visual representation learning: A prospective study. *Machine Intelligence Research*, pages 1–27, 2022.

[25] David Lopez-Paz, Robert Nishihara, Soumith Chintala, Bernhard Scholkopf, and Léon Bottou. Discovering causal signals in images. In *CVPR*, pages 6979–6987, 2017.

[26] Van-Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Grit: Faster and better image captioning transformer using dual visual features. In *ECCV*, pages 167–184, 2022.

[27] Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. *arXiv preprint arXiv:2102.09777*, 2021.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

[29] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[30] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *CVPR*, pages 10860–10869, 2020.

[31] Lin CY ROUGE. A package for automatic evaluation of summaries. In *ACL*, 2004.

[32] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu, Stephen James, Kimin Lee, and Pieter Abbeel. Masked world models for visual control. *arXiv preprint arXiv:2206.14244*, 2022.

[33] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. From show to tell: a survey on deep learning-based image captioning. *PAMI*, 2022.

[34] Jinghan Sun, Dong Wei, Liansheng Wang, and Yefeng Zheng. Lesion guided explainable few weak-shot medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 615–625. Springer, 2022.

[35] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.

[36] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense representation learning via causal inference. In *CVPR Workshops*, pages 378–379, 2020.

[37] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *ICCV*, pages 3091–3100, 2021.

[38] Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. A medical semantic-assisted transformer for radiographic report generation. In *MICCAI*, pages 655–664. Springer, 2022.

[39] 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.

[40] 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 *ICML*, pages 2048–2057. PMLR, 2015.

[41] 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*, page 102510, 2022.

[42] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *PAMI*, 2021.

[43] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *CVPR*, pages 9847–9857, 2021.

[44] 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 *MICCAI*, pages 72–82. Springer, 2021.

[45] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*, 2022.

[46] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021.

[47] S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proceedings of the IEEE*, 109(5):820–838, 2021.

A. VLP Framework

In the pre-training phase, we proposed a novel approach that combines Prefix Language Modeling (PLM) and Masked Image Modeling (MIM) to address the challenges discussed in Section 3.2. The detailed Visual Language Pre-training (VLP) framework is illustrated in Figure 6, which demonstrates its ability to handle diverse data scenarios, such as unpaired and single-modal data.

B. Ablation on Mask Ratio

Masking Ratio	BLEU-1	BLEU-4	CIDEr	Rouge-L
Baseline	0.433	0.148	0.501	0.345
w/ 75%	0.450	0.160	0.486	**0.360**
w/ 85%	**0.452**	**0.161**	**0.522**	0.351
w/ 95%	0.432	0.153	0.460	0.346

Table 5. We evaluated the performance of various masking ratios for MIM on the IU-Xray dataset. In our experiments, we pre-trained the VLP model for 100 epochs and then fine-tuned it in the baseline for an additional 5 epochs.

We conducted ablation experiments to assess the impact of masking ratios on model performance, and the results are presented in Table 5. Our VLP model achieved the best performance with a higher masking ratio of 85%, which is in contrast to the optimal masking ratio of 75% reported by MAE [12]. We attribute this difference to the cross-modal information correlations, where the masked information can be reconstructed by visible features from both language and images. Furthermore, VLP tends to learn general features from the masked modality at higher masking ratios, while distinguishable features can be extracted by the complete information from another modality. To explore whether increasing the masking ratio further would further improve the performance, we experimented with a higher masking ratio of 95%. However, the decreased results in Table 5 indicates that this approach leads to excessive information loss.

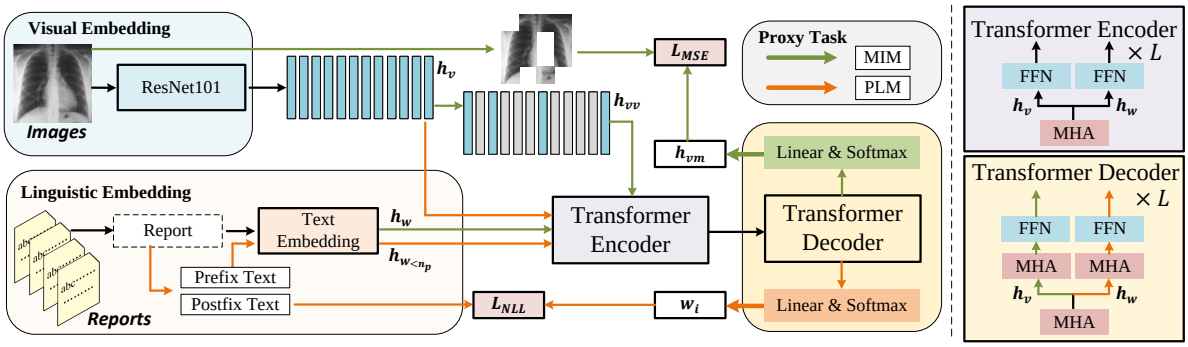


Figure 6. The overview of VLP. MHA and FFN are the Multi-Head Attention layer and Feed-Forward Network layer, respectively.

	Method	Backbone	#Enc	#Dec	$\mathcal{K}$	$\mathcal{T}$	$\mathcal{M}$
Light	R2Gen[3]	Resnet101	3	3			√
	CMCL[20]	Resnet50	/	♠			
	PPKED[21]	Resnet152	2	1	√	√	
	CA[22]	Resnet50	♣	♠		√	
	AlignTransformer[44]	Resnet50	3	3			√
Heavy	M2TR[27]	Densenet151	6	12			√
	RRG-GSK[41]	Resnet101	12	3	√	√	
	MSAT[38]	CLIP	6	6	√		√
	VLCI	Resnet101*	3	3			

Table 6. The details of VLCI and several compared RRG models, the #Enc and #Dec denote the number of transformer layers in the encoder and decoder, respectively. The marker ♣ means 2 Contrastive Attention, and ♠ means Hierarchical LSTM. The backbone of VLCI is the first three blocks of Resnet101. Besides, we show the employed model boosting modules, including the knowledge-aware module $\mathcal{K}$, template retrieval module $\mathcal{T}$, and memory-drive module $\mathcal{M}$.

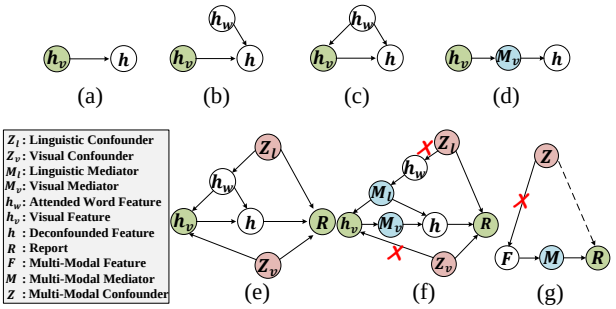


Figure 7. The demonstration of the Structural Causal Model (SCM), (a-d) is the fundamental structure of SCM, and (e-g) is the SCM for the description of VLCI.

C. Model Scale

The RRG models used in our experiment can be categorized as heavyweight or lightweight, as shown in Table 6. Models with no more than 3 layers in both the encoder and decoder are considered lightweight, while others are considered heavyweight. To enhance model performance, these models employ various modules, such as the knowledge-aware module, template retrieval module, and memory-driven memory. In contrast, our lightweight VLCI model only utilizes causal intervention and achieves a significant improvement in performance. Additionally, our VLCI model can be trained on the MIMIC-CXR dataset using only one NVIDIA RTX 3090, whereas the heavyweight MSAT model requires eight NVIDIA TESLA V100 GPUs.

D. Structural Causal Model

To clarify the mechanism of causal intervention, we introduce the Structural Causal Model (SCM) and its symbols, which are shown in Figure 9. The SCM is a mathematical framework used in causal inference to model the relationships between variables and determine cause-and-effect relationships. It uses directed acyclic graphs (DAGs) to represent causal systems, where each variable is a node and the arrows show causal relationships between the variables.

For example, in Figure 9(a), the symbol $h_v \rightarrow h$ represents the visual feature affecting the multi-modal feature. In contrast, the multi-modal feature can be caused by the attended word feature, as shown in Figure 9(b) ($h_v \rightarrow h \leftarrow h_w$). This fork structure indicates that two variables share a common effect or outcome, such as the word "normal" and the visual feature of the heart, leading to the multi-modal feature of "normal heart."

However, the attended word feature can influence the visual feature through cross-attention and cause confounding, as shown in Figure 9(c) ($h_v \leftarrow h_w \rightarrow h \leftarrow h_v$). The word "normal heart" is the linguistic confounder Z_l that affects the visual feature extraction of the pleura through h_w and causes h confounding, as shown in Figure 9(e).

To eliminate this back-door path, we can use causal intervention, which involves changing the value of a variable in the system to observe the resulting changes in other variables. The chain $h_v \rightarrow M_v \rightarrow h$ in Figure 9(d) is a direct

connection from cause to effect, allowing us to manipulate the variable (M_v) affected by the cause variable (h_v) and in turn affecting the effect variable (h). Thus, the multi-modal feature of "small pleural effusion" can be accurately extracted using M_v (the visual feature of "pleural effusion" in each heart size situation). A similar operation can be applied to the linguistic modality to complete the multi-modal front-door intervention, as shown in Figure 9(f, g).

E. Proof of Equation. 5

Assume that $F \triangleq \{h_v, h_w\}$, we can formulate the interventional probability as follows:

$$P(R|do(h_v), do(h_w)) = P(R|do(F)). \tag{11}$$

To eliminate the unobservable confounder, we introduce the mediator M to cut off the link $V \leftarrow Z \rightarrow R$. The total probability $P(R|do(F))$ can be represented as the following summation:

$$\sum_m P(R|do(F), M = m)P(M = m|do(F)), \tag{12}$$

where M is introduced by F without the back-door path. Thus, the intervention probability is equal to the conditional probability in the path $F \rightarrow M$ [23]. Besides, there is no direct causal path between F and R. In this way, the introduced summation in Eq. (12) can be reformulated as:

$$\begin{aligned} &\sum_m P(R|do(F), do(M = m))P(M = m|F) \\ &= \sum_m P(R|do(M = m))P(M = m|F). \end{aligned} \tag{13}$$

To estimate $P(R|do(M = m))$, we can apply the back-door intervention to cut off the link $M \leftarrow F \leftarrow Z \rightarrow R$ [19]. Therefore, we have the intervention probability formulation as follows:

$$\begin{aligned} &P(R|do(M = m)) = \\ &\sum_{\hat{F}} P(R|do(M = m), F = \hat{F})P(F = \hat{F}|do(M = m)) = \\ &\sum_{\hat{F}} P(F = \hat{F})P(R|F = \hat{F}, M = m), \end{aligned} \tag{14}$$

where $\hat{F}$ is the selected features from F not caused by M. At last, via applying Eq. (14), we can further calculate Eq. (13) as follows:

$$\sum_m P(M = m|F) \sum_{\hat{F}} P(F = \hat{F})P(R|F = \hat{F}, M = m). \tag{15}$$

So the proof of Equation. 5 is done.

F. Visualization Result

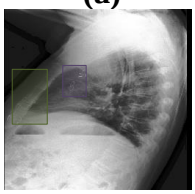
In general, the evaluation of RRG's performance involves three key aspects: keyword detection, determination of whether the identified keywords indicate abnormalities, and generation of lengthy texts. The performance of each aspect can be assessed using specific evaluation metrics. The BELU-1 metric is used to evaluate the accuracy of identifying individual words, while the BELU-4 metric examines the precision of detecting individual abnormalities. On the other hand, the CIDEr metric assesses the overall coherence, logic, and similarity of the entire text. Our approach has yielded noteworthy improvement in both the BELU-1 and BELU-4 metrics. However, our method's performance falls considerably short of knowledge-based models in the CIDEr metric. The inferior performance is attributed to the fact that our method only relies on the report content of the training set, rather than utilizes the medical corpus provided from the knowledge graph. Specifically, the target reports in the training set only include the presence and location of pneumothorax. The limited information (e.g., lack of evidence and description) in the target reports makes it challenging for VLCI to learn such associations.

We further evaluated the errors produced by VLCI by presenting three samples in Figure 8. While our method successfully detected multiple abnormalities, it failed to detect "Fracture" and "Support Device" in Figure 8 (a). This may consistent with our previous statement that single-source abnormalities are challenging to detect. Similarly, Figure 8 (b) shows that VLCI ignored some abnormalities, such as atelectasis and lung opacities, which are difficult to be disentangled from other abnormalities when a patient has multiple abnormalities at the same time. In Figure 8 (c), the ground truth indicated the presence of hydropneumothorax, a condition characterized by the presence of both gas and fluid in the chest, whereas pleural effusion only contains fluid. While VLCI correctly identified the presence of gas and fluid in the chest and proposed pneumothorax, it incorrectly estimated pleural effusion due to the lack of knowledge and incorrect estimation. In this example, our method produced messy text and failed to estimate pneumonia and lung consolidation. Furthermore, while VDM and LDM can help identify visual and language concepts, detecting highly specialized concepts with latent relationships that are not present in the data is difficult.

Abnormality:

①	Lung Opacity	④	Pneumonia	⑦	Pneumothorax	♣	Enlarged Cardiomediastinum
②	Lung Lesion	⑤	Atelectasis	⑧	Pleural Effusion	♠	Cardiomegaly
③	Consolidation	⑥	Edema	⑨	Pleural Other	♥	Fracture
						♦	Support Devices

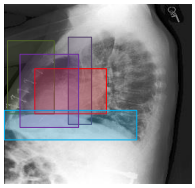
(a)



Ground Truth: Lateral view somewhat limited due to overlying motion artifact. The lungs are low in volume. There is ③ no focal airspace consolidation to suggest pneumonia. A 1.2-cm calcified granuloma just below the medial aspect of the right hemidiaphragm is unchanged from prior study. ⑧No pleural effusions or ⑥ pulmonary edema. There is ⑦no pneumothorax. ♥*The inferior sternotomy wire is fractured but unchanged.* ♦*Surgical clips and vascular markers in the thorax are related to prior CABG surgery.*

VLCI: frontal and lateral views of the chest were obtained . the patient is rotated somewhat to the left. the patient is rotated somewhat to the left. the patient is rotated somewhat to the left. there is ③no focal consolidation ⑧pleural effusion or ⑦ evidence of pneumothorax . ♠the cardiac and ♣mediastinal silhouettes are stable .

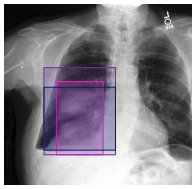
(b)



Ground Truth: ♦*Left-sided pacer is re-demonstrated with leads terminating in the right atrium and right ventricle.* ♥*The patient is status post median sternotomy, aortic valve replacement, and CABG.* ♠*Heart size is mildly enlarged*, unchanged. ♣Mediastinal and hilar contours are similar. Mild upper zone pulmonary vascular redistribution is likely chronic ⑥without overt pulmonary edema. ①*Lung volumes remain low with streaky opacities in the lung bases* ⑤*suggestive of atelectasis.* ⑧No large pleural effusion or ⑦pneumothorax is present. Fusion hardware within the lumbar spine is partially imaged as well as hardware within the right humeral head.

VLCI: frontal and lateral views of the chest were obtained. ♥*the patient is status post sternotomy and cabg.* ♦*left-sided is again seen with leads to the of the right atrium and right ventricle.* the patient is status post sternotomy and cabg. ♠*the cardiac silhouette is mildly enlarged*. the aorta is calcified and tortuous. there is ③no focal consolidation ⑧pleural effusion or ⑦evidence of pneumothorax. ⑥no overt pulmonary edema is seen.

(c)



Ground Truth: ⑦*There is a large right hydropneumothorax with a moderate amount of fluid.* There is no evidence of tension as is supported by the fact that the trachea, the aortic knob, and the left heart border appear in similar position ⑦*as radiograph prior to the pneumothorax on* ___. ①*Hazy opacities are seen involving the right middle and lower lobes.* ④ *The localized nature of this process more likely represents hemorrhage or infectious process* ⑥rather than reexpansion edema. The left lung is clear. ♣The cardiomediastinal silhouette is stable. ♥There are no acute bony abnormalities.

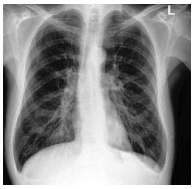
VLCI: frontal and lateral views of the chest were obtained.♦ ***there has been interval removal of a right-sided chest tube*** . there is no significant interval change in the right apical pneumothorax. ⑦*there is a small right apical pneumothorax .* ⑧***there is a small right pleural effusion***. there is a small amount of subcutaneous emphysema in the right chest wall . there is a small amount of subcutaneous emphysema in the right lateral chest wall. there is a small amount of subcutaneous emphysema. there is a small amount of subcutaneous emphysema in the right chest wall.

Figure 8. The erroneous result of our VLCI models on the MIMIC-CXR dataset is presented. Thirteen kinds of abnormalities are marked with different markers and colors. Note that keywords in the reports are also marked with different markers and colors. Correctly identified abnormalities are marked in the corresponding color, while other descriptions in bold, italics, and underscores are incorrect. Descriptions marked only with underscores indicate repeated words.

Abnormality:

① Lung Opacity ④ Pneumonia ⑦ Pneumothorax ♣ Enlarged Cardiomediastinum
② Lung Lesion ⑤ Atelectasis ⑧ Pleural Effusion ♥ Cardiomegaly
③ Consolidation ⑥ Edema ⑨ Pleural Other ♦ Fracture
♠ Support Devices

No Finding

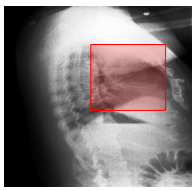


Ground Truth: Right-sided Port-A-Cath terminates in the mid SVC as before. ♠Heart is top-normal in size. ♣Mediastinal and hilar contours are within normal limits. Lung volumes are low over the lungs are clear without ③focal consolidation, ⑧effusion or ⑦pneumothorax.

Baseline: as compared to the previous radiograph there is no relevant change. low lung volumes. ♠borderline size of the cardiac silhouette without pulmonary edema. ⑧no pleural effusions. ④no pneumonia.

VLCI: frontal and lateral views of the chest were obtained. the patient is status post sternotomy and cabg. ♠the heart size is normal. ♣the mediastinal and hilar contours are normal. the pulmonary vasculature is normal. lungs are clear. ⑧no pleural effusion or ⑦pneumothorax is seen. ♥ there are no acute osseous abnormalities.

♠

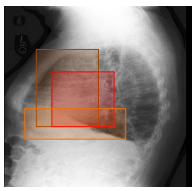


Ground Truth: there is no evidence of acute cardiopulmonary disease. ④No pneumonia, vascular congestion, or ⑧ pleural effusion. ♠*he cardiac silhouette is at the upper limits of normal in size or slightly enlarged.*

Baseline: as compared to the previous radiograph there is no relevant change. low lung volumes. ♠*borderline size of the cardiac silhouette* ⑥without pulmonary edema. ⑧no pleural effusions. ④no pneumonia.

VLCI: frontal and lateral views of the chest were obtained. lung volumes are low in bronchovascular crowding. there is ③ no focal consolidation ⑧pleural effusion or ⑦pneumothorax. ♠*the cardiac silhouette is mildly enlarged.* ♣mediastinal and hilar contours are normal.

⑥⑧♠

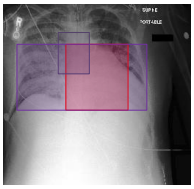


Ground Truth: Interstitial prominence has increased compared to prior, ⑥*suggestive of mild edema.* ③No focal consolidation or ⑦pneumothorax is detected. ⑧*Tiny right pleural effusion appears* new compared to prior. Heart and mediastinal contours appear stable ♠*with mild cardiomegaly.*

Baseline: frontal and lateral views of the chest were obtained. ⑥*there is diffuse increase in interstitial markings bilaterally which may be due to mild interstitial edema* although atypical infection is not excluded in the appropriate clinical setting. ⑧ *no large pleural effusion is seen*. ⑦there is no pneumothorax. ♠the cardiac and ♣mediastinal silhouettes are stable.

VLCI: ap and lateral views of the chest. ⑥*there is mild pulmonary edema* with ⑧*small bilateral pleural effusions*. ♠*the heart is mildly enlarged*. the mediastinal contour is stable with atherosclerotic calcification along the aortic. ♥bony structures are intact

♦♠①

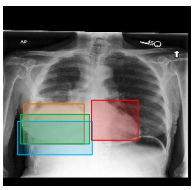


Ground Truth: ♦*The ET tube is 3.5 cm above the carina.* Right IJ Cordis tip is in the proximal SVC. ♠*The heart size is moderately enlarged.* ① *There is ill-defined vasculature and alveolar infiltrate*, right greater than left. This is markedly increased compared to the film from two hours prior and likely represents fluid overload.

Baseline: ♦*endotracheal tube tip terminates approximately 45 cm from the carina.* ♠*heart size is mildly enlarged.* ♣mediastinal contours are unremarkable . there is crowding of the bronchovascular structures ⑥***with mild pulmonary edema .*** patchy opacities in the lung bases ⑤***may reflect areas of atelectasis.*** ⑧no large pleural effusion or ⑦ pneumothorax is identified . ♥no acute osseous abnormalities are seen .

VLCI: ♦*the tip of the tube projects _ cm above the carina.* the tube is in position. ⑦there is no evidence of no pneumothorax. the lung volumes are low. ①*there are diffuse bilateral airspace opacities with air.* ♠*the cardiac silhouette remains enlarged.*

⑤③⑧

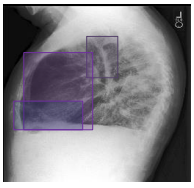


Ground Truth: Chest PA and lateral radiograph demonstrates a markedly elevated right hemidiaphragm with adjacent compressive ⑤*atelectasis* or ③*consolidation*. Minimal blunting of the posterior costophrenic angle ⑧*may indicate a small right pleural effusion*. Left lung is clear. ♣Cardiomediastinal borders are unremarkable.

Baseline: frontal and lateral views of the chest were obtained. there is persistent elevation of the right hemidiaphragm ⑤*with overlying atelectasis.* ③***no definite focal consolidation is seen***. ⑧***there is no pleural effusion*** or ⑦pneumothorax. the cardiac and mediastinal silhouettes are stable.

VLCI: frontal and lateral views of the chest were obtained. ⑧*there is a small right pleural effusion* ⑤*with overlying atelectasis* ③*underlying consolidation is not excluded*. the left lung is clear. ♠the cardiac silhouette is not enlarged. the aorta is calcified and tortuous. ⑦no pneumothorax is seen.

♦①④



Ground Truth: ♦*Esophageal stent is again seen*, appears more inferior in position as compared to the prior study. Right perihilar chronic changes are seen. ①*There is slight increase in the right mid lung opacity* which ④*could be due to underlying infection, possibly in the superior right lower lobes*. ⑦No pneumothorax is seen.

Baseline: there is diffuse increase in interstitial markings bilaterally which ⑥***may be due to mild interstitial edema.*** ⑧no large pleural effusion is seen. ⑦there is no pneumothorax. ♠the cardiac and ♣mediastinal silhouettes are stable.

VLCI: ①*there has been interval development of a right upper lobe opacity* ④*which is for pneumonia*. ⑧***there is also a small right pleural effusion.*** ⑦there is no pneumothorax. ♣the cardiomediastinal silhouette is within normal limits. ♥no acute osseous abnormalities identified .

Figure 9. The results of the Baseline and VLCI models on the MIMIC-CXR dataset are presented. Thirteen kinds of abnormalities are marked with different markers and colors. Note that keywords in the reports are also marked with different markers and colors. Correctly identified abnormalities are marked in the corresponding color, while other descriptions in bold, italics, and underscores are incorrect.