

CmEAA: Cross-modal Enhancement and Alignment Adapter for Radiology Report Generation

Xiyang Huang¹, Yingjie Han¹, Yaoxu Li¹, Runzhi Li²,
Pengcheng Wu¹, Kunli Zhang^{1*}

¹School of Computer Science and Artificial Intelligence, Zhengzhou University

²Cooperative Innovation Center of Internet Healthcare, Zhengzhou University

iek1zhang@zzu.edu.cn

Abstract

Automatic radiology report generation is pivotal in reducing the workload of radiologists, while simultaneously improving diagnostic accuracy and operational efficiency. Current methods face significant challenges, including the effective alignment of medical visual features with textual features and the mitigation of data bias. In this paper, we propose a method for radiology report generation that utilizes a **Cross-modal Enhancement and Alignment Adapter (CmEAA)** to connect a vision encoder with a frozen large language model. Specifically, we introduce two novel modules within CmEAA: Cross-modal Feature Enhancement (CFE) and Neural Mutual Information Aligner (NMIA). CFE extracts observation-related contextual features to enhance the visual features of lesions and abnormal regions in radiology images through a cross-modal enhancement Transformer. NMIA maximizes neural mutual information between visual and textual representations within a low-dimensional alignment embedding space during training and provides potential global alignment visual representations during inference. Additionally, a weights generator is designed to enable the dynamic adaptation of cross-modal enhanced features and vanilla visual features. Experimental results on two prevailing datasets, namely, IU X-Ray and MIMIC-CXR, demonstrate that the proposed model outperforms previous state-of-the-art methods.

1 Introduction

Radiological imaging, such as chest X-rays, plays an indispensable role in clinical diagnosis and treatment (Lambin et al., 2017). For radiologists, the process of interpreting radiology images and drafting reports is both time-intensive and susceptible to errors (Chen et al., 2020). The objective of automatic radiology report generation (RRG) is to

* Corresponding author

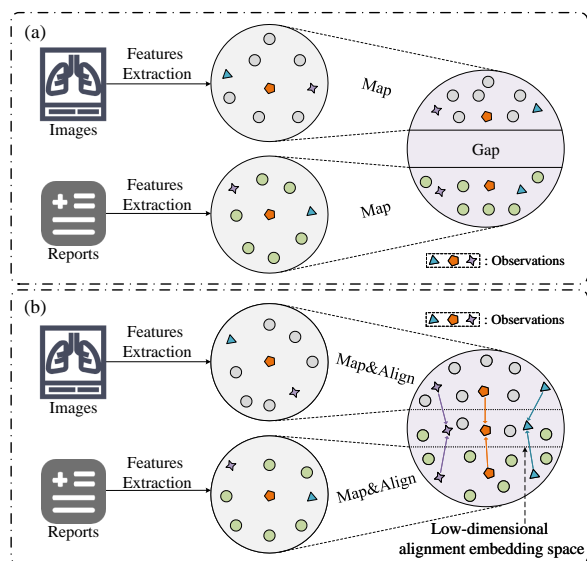


Figure 1: Comparison of two alignment methods.(a) illustrates that in a unified embedding space, it is challenging to align relevant features due to the inherent gap between different modalities. (b) showcases our alignment method within a low-dimensional embedding space, achieving more effective modalities alignment in contrast to (a).

synthesize natural language descriptions from medical images that delineate observed abnormalities and lesions (Goergen et al., 2013). This task aims to alleviate the burden on radiologists, particularly in regions with limited medical resources. Most existing methods for RRG (Yuan et al., 2019; Xue et al., 2018) are typically refinements derived from image captioning (Cornia et al., 2020), leading to significant advancements. Despite the remarkable performance, these methods suffer from such data bias: the normal cases dominate the dataset over the abnormal cases (Huang et al., 2023). This data bias causes the model to primarily learn representations of normal samples during training, resulting in inadequate detection of abnormal lesions in the images. Furthermore, cross-modal mapping between images and text is crucial for generating high-quality

reports. Due to the absence of annotated correspondences between images and text, directly aligning them remains challenging. As shown in Figure 1(a), some methods involve mapping the visual features extracted from images and the textual embeddings derived from ground truth reports into a unified embedding space. Nonetheless, this straightforward mapping of representations fails to bridge the inherent gap between image features and text features, leading to suboptimal alignment. As depicted in Figure 1(b), our proposed method aligns image and text features within a low-dimensional alignment embedding space, which differs to previous methods(Chen et al., 2022), effectively alleviating this issue.

To tackle these challenges, we introduce a novel approach for RRG that employs a Cross-modal Enhancement and Alignment Adapter (CmEAA) to connect a vision encoder with a frozen large language model. Within CmEAA, the Cross-modal Feature Enhancement (CFE) is designed to enhance the visual features with diverse radiological semantic information. Specifically, we obtain observation labels according to the probabilities of X-ray images on 14 types of radiological observations. Based on observation labels, high co-occurrence observation-related n-grams are identified by calculating the pointwise mutual information(Church and Hanks, 1990) between various observations and n-grams derived from the training set. The observation contextual embeddings, derived by embedding observation-related n-grams, are used to enhance the visual features of lesions and abnormal regions in radiology images via a cross-modal enhancement Transformer. The enhanced visual features are then adapted to vanilla visual features via a set of dynamic weights. To facilitate the global alignment of images and text representations, we develop a Neural Mutual Information Aligner (NMIA) that maximizes mutual information between images and text representations within a low-dimensional alignment embedding space during training and provides potential global alignment visual representations through two simple linear layers during inference. Our contributions can be summarized as follows:

- To facilitate the enhancement of visual features and global cross-modal alignments, we propose a novel Cross-modal Enhancement and Alignment Adapter with two modules: Cross-modal Feature Enhancement (CFE) and Neural Mutual

Information Aligner (NMIA).

- CFE performs cross-modal feature enhancement to visual features, enabling the model to learn hidden radiological visual representations that capture comprehensive semantic information pertaining to lesions and abnormal regions. NMIA accomplishes global alignment of visual and text representations through the maximization of neural mutual information.
- We evaluate our method on two public radiology report generation datasets, IU-Xray(Demner-Fushman et al., 2016) and MIMIC-CXR(Johnson et al., 2019). The experimental results demonstrate the effectiveness of our method and we also conduct a detailed case analysis to illustrate the benefits of CFE and NMIA.

2 Methods

2.1 Overview

As shown in Figure 2, our proposed CmEAA is built upon two RepAdapters(Luo et al., 2023) and contains two novel modules: Cross-modal Feature Enhancement (CFE) and Neural Mutual Information Aligner(NMIA). During training, CFE leverages the observation-related contextual embeddings and cross-modal enhancement Transformer to conduct cross-modal feature enhancement on visual features, as detailed in section 2.2. NMIA maps image and text representations to the same low-dimensional alignment embedding space and then aligns the cross-modal representations by maximizing the neural mutual information between both with the multimodal representations, which will be introduced in section 2.3.

Given a chest X-ray image X , we use Swin Transformer(Liu et al., 2021) as the visual features extractor to extract the vanilla visual features $X_i = \{x_1, x_2, \dots, x_i\}$. The entropy of features across different modalities varies, indicating diverse levels of information content. In light of this, we introduce a set of dynamic weights to achieve adaptation between the cross-modal enhanced features and the vanilla visual features. For the visual features X_i , we first apply average pooling to obtain the pooled feature X_p , which is used to generate dynamic weights. The weights generator in the Weight&Fusion module consists of a linear layer and a Softmax function, which can be formulated as:

$$X_p = Pooling(X_i) \quad (1)$$

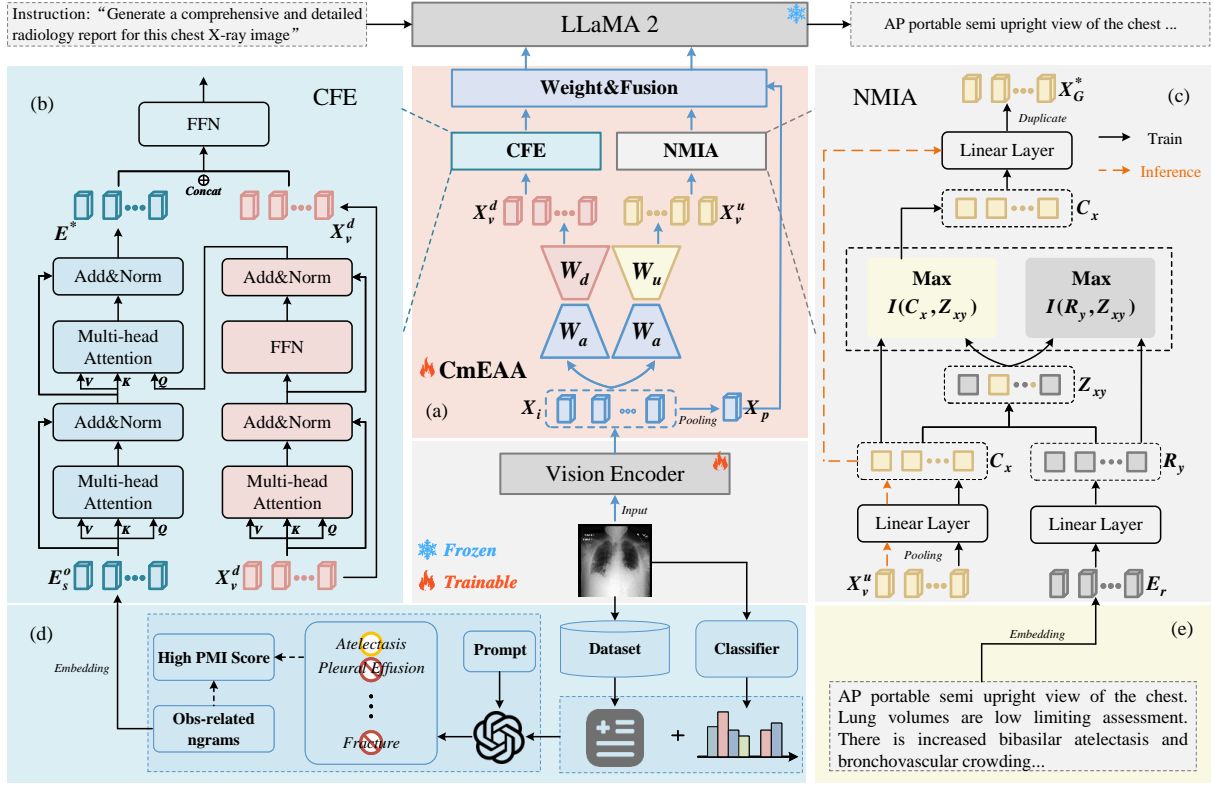


Figure 2: An overview of our proposed method. The visual encoder and frozen large language model are connected by CmEAA, which includes two main modules: Cross-modal Feature Enhancement (CFE) and Neural Mutual Information Aligner (NMIA).

$$\hat{w} = \text{Softmax} \left(\frac{\text{Linear}(X_p)}{\tau} \right) \quad (2)$$

where τ is the temperature of the Softmax. Thus, the CmEAA can be defined by:

$$\tilde{X}^* = \text{CmEAA}(X_i) \quad (3)$$

$$\text{CmEAA}(X_i) = \hat{w}_0 \cdot \text{CFE}(f_d(X_i)) + \hat{w}_1 \cdot f_u(X_i) + \text{NMIA}(f_u(X_i)) \quad (4)$$

Here, f_d and f_u are RepAdapters(Luo et al., 2023). \hat{w}_0 and \hat{w}_1 are weights derived from Eqn.2. The downsampling projection of two adapters are shared. Following the application of CmEAA to X_i , the visual representations \tilde{X}^* are derived. Subsequently, \tilde{X}^* are concatenated with the text instruction X_P and fed into the frozen Llama2 model for decoding and generating the final report X_r . The basic training objective of our method for language modeling is defined as:

$$\mathcal{L}_{CE} = - \sum_{i=1}^L \log p_{\psi} \left(x_i | \tilde{X}^*, X_P, X_{r, < i} \right) \quad (5)$$

where ψ is the learnable parameters. $X_{r, < i}$ is the report tokens before the current prediction token X_i .

2.2 Cross-modal Feature Enhancement

Relying exclusively on image features to generate corresponding text presents challenges in maintaining consistency between images and detailed text reports. Nonetheless, excessively complex external knowledge may interfere with the model’s attention distribution, resulting in alterations to the learned representations. Accordingly, we utilize precise and well-curated observation-related semantic information from the training set to facilitate cross-modal enhancement. In CheXpert(Irvin et al., 2019), 14 observations are defined for X-ray images, encapsulating the key visual information present in the images. Each observation label is classified as Present, Absent, or Uncertain. Observations encapsulate a high-level synthesis of the radiological image, whereas the finalized generated report must encompass more granular diagnostic insights. Consequently, it is important to strengthen the connection between lesions, abnormal regions, and their pertinent contextual information, while ensuring precise lesion prediction. Therefore, leveraging these fine-grained observation-related contextual information for cross-modal enhancement of visual features could mitigate the impact of the

Algorithm 1 \mathcal{L}_{NMI} Minimization with MINE

$\theta, \phi \leftarrow$ initialize network parameters

for each training iteration **do**

Draw b minibatch samples from the joint distribution:

$$(C_x^{(1)}, Z_{xy}^{(1)}), \dots, (C_x^{(b)}, Z_{xy}^{(b)}) \sim \mathbb{P}_{C_x Z_{xy}}$$

$$(R_y^{(1)}, Z_{xy}^{(1)}), \dots, (R_y^{(b)}, Z_{xy}^{(b)}) \sim \mathbb{P}_{R_y Z_{xy}}$$

Draw n samples from the Z_{xy} marginal distribution:

$$(\bar{Z}_{xy}^{(1)}, \dots, \bar{Z}_{xy}^{(b)}) \sim \mathbb{P}_{Z_{xy}}$$

Evaluate the lower-bound:

$$I_c(C_x; Z_{xy}) = \frac{1}{b} \sum_{i=1}^b T_\theta(C_x^{(i)}, Z_{xy}^{(i)}) - \log\left(\frac{1}{b} \sum_{i=1}^b e^{T_\theta(C_x^{(i)}, \bar{Z}_{xy}^{(i)})}\right)$$

$$I_r(R_y; Z_{xy}) = \frac{1}{b} \sum_{i=1}^b T_\theta(R_y^{(i)}, Z_{xy}^{(i)}) - \log\left(\frac{1}{b} \sum_{i=1}^b e^{T_\theta(R_y^{(i)}, \bar{Z}_{xy}^{(i)})}\right)$$

Update θ, ϕ by minimizing:

$$\mathcal{L}_{NMI}(\theta, \phi) = -(I_c + I_r)$$

end for

aforementioned challenges.

As shown in Figure 2(d), to extract the observation labels of a given image, we adapt the `cft-chexpert`¹ to obtain the observation probabilities within 14 categories $C = \{C_1, C_2, \dots, C_{14}\}$ as indicated in (Irvin et al., 2019). We then use the observation probabilities and the corresponding report of the image to prompt GPT-4 to obtain the observation labels, with specific details provided in Appendix A.1. Following (Hou et al., 2023b), we obtain observation-related n-grams as contextual information relevant to visual features. Given a predefined observation set $O = \{o_1, o_2, \dots, o_n\}$ and n-gram units $S = \{s_1, s_2, \dots, s_t\}$ based on the training reports, Organ(Hou et al., 2023b) extracts the observation-related n-gram units with $PMI(o_i, s_j)$:

$$PMI(o_i, s_j) = \log \frac{p(o_i, s_j)}{p(o_i)p(s_j)} \quad (6)$$

where o_i is the i -th observation, s_j is the j -th n-gram, and $p(o_i, s_j)$ is the frequency that an n-gram s_j appears in a report with observation o_i in the training set. A higher PMI score implies two units with higher co-occurrence. Then, a set of observation-related n-grams $S_k^o = \{s_1^o, s_2^o, \dots, s_k^o\}$ is derived and the observation-related contextual embeddings E_s^o are obtained by embedding S_k^o through Llama2. As shown in Figure 2(b), cross-modal enhancement Transformer in CFE consists

¹<https://github.com/maxium0526/cft-chexpert>

of two Transformer(Vaswani, 2017) submodules:

(1) A text Transformer performs self-attention for observation-related contextual embeddings E_s^o and interacts with visual features through cross-attention. (2) A visual Transformer that performs self-attention on visual features X_v^d and transmits them as queries to the text Transformer. The concatenated outputs of the text Transformer E^* and visual features X_v^d are processed through a FFN layer to yield the enhanced features. The process is formally defined as follows:

$$X'_v = CFE(E_s^o, X_v^d) \quad (7)$$

$$CFE(E_s^o, X_v^d) = FFN(E^* \oplus X_v^d) \quad (8)$$

$$E^* = CA(X_{sa}^v, E_{sa}^s, E_{sa}^s) \quad (9)$$

$$E_{sa}^s = SA(E_s^o), X_{sa}^v = SA(X_v^d) \quad (10)$$

where X'_v are the cross-modal enhanced features. CA and SA represent cross-attention and self-attention, respectively. X_v^d are visual features processed by f_d .

2.3 Neural Mutual Information Aligner

The cross-modal alignment strategies employed in RRG task can be divided into two categories: (1) aligning visual features with abnormality(pathologic) labels, and (2) executing global alignment between visual and textual representations. Nevertheless, the first approach is limited by the paucity of annotated data and the precision of lesion classification. Furthermore, labels represent a high-dimensional abstraction of visual data and provide relatively sparse information. Consequently, we introduce NMIA to globally align visual and textual representations under a neural mutual information loss. Mutual information quantifies the dependence of two random variables X and Y . In contrast to correlation, mutual information captures non-linear statistical dependencies between variables(Belghazi et al., 2018), and thus can act as a measure of true dependence. However, only in a few special cases can one calculate the exact value of mutual information(Cheng et al., 2020), since the calculation requires closed forms of density functions and a tractable log-density ratio between the joint and marginal distributions. Therefore, we employ the Mutual Information Neural Estimator(Belghazi et al., 2018) to approximate

Dataset	Model	NLG Metrics					CE Metrics		
		BL-1	BL-2	BL-3	BL-4	RG-L	P	R	F1
IU X-Ray	R2Gen	0.470	0.304	0.219	0.165	0.371	-	-	-
	R2GenCMN	0.475	0.309	0.222	0.170	0.375	-	-	-
	M2KT	<u>0.497</u>	0.319	<u>0.230</u>	<u>0.174</u>	0.399	-	-	-
	METransformer	0.483	<u>0.322</u>	0.228	0.172	0.380	-	-	-
	MAN	0.501	0.328	<u>0.230</u>	0.170	0.386	-	-	-
	XrayGPT(7B)	0.177	0.104	0.047	0.007	0.203	-	-	-
	Ours	0.481	0.319	0.234	0.181	<u>0.392</u>	-	-	-
MIMIC -CXR	R2Gen	0.353	0.218	0.145	0.103	0.277	0.333	0.273	0.276
	R2GenCMN	0.353	0.218	0.148	0.106	0.278	0.334	0.275	0.278
	M2KT	0.386	0.237	0.157	0.111	0.274	<u>0.420</u>	<u>0.339</u>	0.352
	METransformer	0.386	<u>0.250</u>	<u>0.169</u>	<u>0.124</u>	0.291	0.364	0.309	0.311
	MAN	<u>0.396</u>	0.244	<u>0.162</u>	0.115	0.274	0.411	0.398	<u>0.389</u>
	XrayGPT(7B)	0.128	0.045	0.014	0.004	0.079	-	-	-
	Med-PaLM(562B)	0.317	-	-	0.115	0.275	-	-	-
	Ours	0.407	0.255	0.174	0.126	<u>0.281</u>	0.505	0.330	0.399

Table 1: Performance comparisons of the proposed method with existing methods on the test sets of MIMIC-CXR and IU-Xray with respect to NLG and CE metrics. The best results are highlighted in **bold**, while the second-best results are underlined.

the lower bound of mutual information between images and text representations. MINE uses a lower-bound to the MI based on the Donsker-Varadhan representation (Donsker and Varadhan, 1983) of the KL-divergence, which is formulated as:

$$D_{KL}(\mathbb{P}||\mathbb{Q}) = \sup_{T:\Omega \rightarrow \mathbb{R}} \mathbb{E}_{\mathbb{P}}[T] - \log(\mathbb{E}_{\mathbb{Q}}[e^T]) \quad (11)$$

Then, MINE is defined as:

$$I(\widehat{X}; Z)_n = \sup_{\theta \in \Theta} \mathbb{E}_{\mathbb{P}_{XZ}^{(n)}} [T_{\theta}] - \log(\mathbb{E}_{\mathbb{P}_X^{(n)} \otimes \mathbb{P}_Z^{(n)}} [e^{T_{\theta}}]) \quad (12)$$

where T_{θ} is a discriminator function modeled by a neural network with parameters θ . In practice, images and text representations are regarded as random variables, denoted as C_x and R_y , respectively. As shown in Figure 2(c), we apply average pooling to the visual features. The Llama2 tokenizer is employed to process the ground truth report text. Both types of features are mapped to an alignment embedding space $\mathbb{R}^{b \times 1 \times k}$ using linear layers. The inherent disparity between visual and textual features may lead to training instability when directly maximizing the neural mutual information between images and text representations, potentially hindering the model’s convergence. Consequently, we obtain the fused features Z_{xy} by summing the two types of features and then separately maximize the neural mutual information between the images features and the fused features, as well as between the

text features and the fused features during training. Details on the implementation of NMIA are provided in Algorithm 1. Following the computation of neural mutual information, an additional linear layer is utilized to project the visual features into the same embedding space with X'_v as the globally aligned visual representations. During training, NMIA calculates the neural mutual information loss in a low-dimensional alignment space, maximizing the mutual information by updating the parameters of the linear layers. During the inference phase, the same trained linear layers are used to obtain the image representations aligned with the text representations. The process is formally defined as follows:

$$X_G^* = NMIA(X_v^u) \quad (13)$$

$$NMIA(X_v^u) = Concat(X_{g(1)}^*, X_{g(2)}^*, \dots, X_{g(n)}^*) \quad (14)$$

$$X_g^* = W_2 \cdot (W_1 \cdot X_v^u + b_1) + b_2 \quad (15)$$

where $X_{g(i)}^*$, $i \in [1, n]$ is a replica of X_g^* . W_1 and W_2 are the weights of the dimensionality reduction and dimensionality expansion linear layers, respectively. As stated in Eqn.14 and Eqn.4, across both phases, the globally aligned features are replicated n^2 times and subsequently incorporated into the

² n is the number of visual tokens

output of the CFE, thereby offering global alignment features for each visual token. The overall loss function of our approach is formulated as:

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda \mathcal{L}_{NMI} \quad (16)$$

where λ is the scale factor.

3 Experiment

This section provides an overview of the experimental datasets and the evaluation metrics employed. The implementation details can be found in Appendix A.3.

3.1 Configurations

Datasets We evaluate our model on two widely-used benchmarks for RRG: IU X-Ray (Demner-Fushman et al., 2016) and MIMIC-CXR (Johnson et al., 2019). IU X-Ray from Indiana University is a relatively small but publicly available dataset containing 7,470 chest X-ray images and 3,955 radiology reports. We split the dataset into train/validation/test sets with a ratio of 7:1:2, which is the same data split as in (Chen et al., 2020). MIMIC-CXR provided by the Beth Israel Deaconess Medical Center. The dataset consists of 377,110 chest X-ray images and 227,835 reports. We adopt the standard train/validation/test splits. The statistics of the datasets can be found in Appendix A.2.

Metrics Following previous research (Chen et al., 2020), we employ the widely-used natural language generation (NLG) metrics and clinical efficacy (CE) metrics. The NLG metrics include BLEU (Papineni et al., 2002) and ROUGE-L (Lin, 2004). And for the clinical efficacy, we apply CheXpert (Irvin et al., 2019) for MIMIC-CXR dataset to label the generated reports with 14 categories. According to these annotations, the Precision, Recall and F1 scores can be calculated over the generated reports and the ground truth reports as the Clinical Efficacy (CE) metrics. Higher is better for all metrics.

3.2 Comparison with State-of-the-Arts

To demonstrate the effectiveness, we compare the performances of our model with a wide range of state-of-the-art models on the MIMIC-CXR and IU X-Ray. The models we compare to include R2Gen (Chen et al., 2020), R2GenCMN (Chen et al., 2022), M2KT (Yang et al., 2023), MAN (Shen et al., 2024), XrayGPT (Thawakar et al., 2024), et al. Table 1 shows the comparison results on NLG and

Model	CFE	NMIA	IU X-Ray		
			BL-1	BL-4	RG-L
(a)	-	-	0.467	0.168	0.376
(b)	✓		0.468	0.172	0.380
(c)		✓	0.479	0.170	0.393
(d)	✓	✓	0.481	0.181	0.392

Table 2: Ablation results of our model and its variants, where the model (a) replaces CmEAA with a linear layer.

k	BL-1	BL-4	RG-L	Δ
32	0.464	0.175	0.378	+0.1%
64	0.458	0.176	0.385	+0.5%
128	0.465	0.180	0.379	+1.3%
256	0.481	0.181	0.392	+3.7%
512	0.465	0.172	0.374	-0.7%

Table 3: Models with different dimensions of the low-dimensional embedding space. The Δ in the table refers to the average variation relative to model (b).

CE metrics. Our method outperforms most of the baselines and achieves state-of-the-art performance. Specifically, on the IU X-Ray dataset, our method achieves the best results on BLEU-3 and BLEU-4. On the MIMIC-CXR dataset, our method achieves the best results on all BLEU metrics and the second-best result on Rouge-L. This indicates that by introducing the CmEAA, our model can generate more coherent reports than baselines. Nonetheless, on IU X-Ray dataset, we notice that our model still exhibits a performance disparity when compared to the best baseline (i.e., MAN (Shen et al., 2024)) on BLEU-1 and BLEU-2. This discrepancy could stem from two factors: (1) Given that the reports in IU X-Ray are relatively brief, n-grams of lengths 1 and 2 become predominant among the extracted n-grams.³ (2) The classifier was not specifically trained on IU X-Ray, resulting in the prediction of incorrect observation labels, which further led to the extraction of erroneous n-grams. Nevertheless, the introduction of globally aligned visual representations ensures that the model maintains robust performance on BLEU-3 and BLEU-4 metrics, demonstrating the benefits of cross-modal global alignment within the NMIA. Experiment results on MIMIC-CXR also validate this finding. For the Clinical Efficacy (CE) metrics, our model achieves the best results on Precision and

³The statistics of observation-related n-grams with different lengths can be found in Appendix A.2


Radiology Image	Ground Truth	Report
 <p> Atelectasis_True Pleural Effusion_False Support Devices_True : Fracture_False </p>	<p>A single portable ap semi-upright view of the chest was obtained. Right ij central venous catheter projects over the right atrium . An icd pacing device with biventricular leads appears unchanged in position . Lung volumes remain low with right basilar atelectasis . Cardiomeastinal silhouette is stable . There is no focal consolidation or pleural effusion . no pneumothorax .</p>	<p>Baseline Ap portable upright view of the chest . Left chest wall pacer device is noted with leads terminating in the right atrium and right ventricle . There is mild pulmonary vascular congestion without overt pulmonary edema . No large pleural effusion or pneumothorax is seen . Cardiomeastinal silhouette is unchanged . Bony structures are intact .</p> <p>CmEAA Portable semi-erect chest radiograph demonstrates low lung volumes with bibasilar atelectasis . There is no pneumothorax or pleural effusion . The cardiomeastinal silhouette is within normal limits . A right internal jugular central venous catheter terminates in the right atrium . Left internal jugular catheter terminates in the right atrium .</p>

Figure 3: An illustration of the report generated by Baseline and the proposed method. The Baseline corresponds to model (a) introduced in Section 3.3. Below the image are observation labels. The contents of the generated report that correspond to the ground truth report and observation labels are highlighted in the same color(blue and green). The areas marked in red correspond to the model’s incorrect predictions.

F1. Specifically, our model achieves 0.505 precision and 0.399 F1, with up to 20.2% and 2.5% compared to the best baseline. In addition, 0.330 recall is achieved by CmEAA, which is competitive result. The CE results indicate that our model can successfully maintain the clinical consistency between the images and the reports.

3.3 Ablation Study

To verify the effectiveness of CmEAA, we do ablation study on IU X-Ray dataset, which is shown in Table 2. The visual encoder and large language model are preserved unchanged, with a simple linear layer employed to replace CmEAA as the model (a). There are two variants: (b) represents the model containing only the CFE module within CmEAA, and (c) depicts the model incorporating solely the NMIA module within CmEAA. Specifically, in comparison to (a), (d) achieves an average improvement of 4.9% across the BLEU-1, BLEU-4, and ROUGE-L metrics, demonstrating the effectiveness of CmEAA in improving the model’s long-text generation capabilities. In comparison to the full model (d), the performance of (b) and (c) significantly deteriorates. This indicates that CFE and NMIA play a vital role in generating reports. Notably, (c) improves on BLEU-1 and Rouge-L by 2.5% and 4.5% compared to (a), while (b) shows limited improvement on BLEU-1 metric. This result is also consistent with previous finding that the efficacy of CFE is constrained by the performance of the classifier, whereas NMIA demonstrates greater stability.

Furthermore, we conduct experiments in differ-

ent dimensions (k) of the low-dimensional alignment embedding space from 32 to 512 on IU X-Ray to study its impact on the results. As shown in Table 3, the model attains optimal performance when k is set to 256. When k is too small (e.g., 32 or 64), the model exhibits only modest improvements. This limitation is likely attributable to the reduced dimensionality of the embedding space, which inadequately accommodates cross-modal alignment information. Consequently, this information may be compressed or lost, resulting in suboptimal alignment of the relevant features. On the other hand, relatively large k can have adverse effects(e.g., 512), potentially due to overfitting. This does not imply that the value of k has a decisive impact on the model’s performance. However, choosing an appropriate value for k is advantageous for attaining improved results.

3.4 Qualitative Analysis

To further understand the effectiveness of our model, qualitative examples are given in Figure 3. Intuitively, the report generated by our model are both accurate and robust, which shows significant alignment with ground truth reports. As the Figure 3 shows, the report generated by our model closely align with the corresponding observation labels and accurately identify the abnormal finding "low lung volumes", the support device "venous catheter", as well as the normal finding "cardiomeastinal silhouette is within normal limits". In contrast, the baseline failed to detect the abnormal finding "low lung volumes", and it incorrectly identifies the support device "venous catheter" as "pacer device". It

is pertinent to highlight that "*low lung volumes*" may indicate the presence of a pathologic condition in the lungs, such as "*Atelectasis*". The term "*low lung volumes*" was not defined as an observation in CheXpert(Irvin et al., 2019). Considering that the X-ray image illustrated in Figure 3 was labeled with the Observation "*Atelectasis true*", the n-grams related to this label encompass pertinent descriptions related to "*low lung volumes*". When observation contextual embeddings were used to enhance the visual features in CFE, semantic information related to "*low lung volumes*" achieved an effective adaptation to visual features. Consequently, our model correctly generated this finding, while the baseline failed to report it in its result. This further illustrates the effectiveness of CmEAA in bridging the gap between visual and textual features.

4 Related work

4.1 Radiology Report Generation

RRG is an application of image captioning methods in the medical field. Compared to image captioning, RRG not only puts forward higher requirements on the length of generated reports but also presents greater challenges on the accuracy of long contextual descriptions(Huang et al., 2023). With the good performance of Transformer(Vaswani, 2017) in various vision and language tasks, a plethora of Transformer-based methods have been explored to enhance image description performance. R2Gen(Chen et al., 2020) records information about previous generation processes through relational memory (RM), where similar patterns in different radiology reports can be implicitly modeled and remembered during the generation process to facilitate report generation. RECAP(Hou et al., 2023a) can capture both spatial and temporal information for generating precise and accurate free-text reports. Additionally, several studies emphasize improving cross-modal alignments in RRG. R2GenCMN(Chen et al., 2022) utilizes a memory matrix as an intermediary between the visual and textual modalities to strengthen global cross-modal alignments. M2KT(Yang et al., 2023) introduces a multimodal alignment module that aligns the pooled visual features with the pooled textual representations derived from a BERT model, alongside the predicted pathologic condition labels. Additionally, some works(Zhang et al., 2020; Huang et al., 2023) began to explore enhancing report generation with additional knowledge. (Zhang et al.,

2020) utilized a pre-constructed graph embedding module (modeled with a graph convolutional neural network) on multiple disease findings to assist the generation of reports. Kiut(Huang et al., 2023) injects clinical knowledge by constructing a symptom graph, combining it with the visual and contextual information, and distilling them when generating the final words in the decoding stage.

4.2 Vision-Language Models

The convergence of computer vision and natural language processing has facilitated the emergence of vision-language models, which integrate visual and linguistic models to attain cross-modal comprehension and inferential abilities(Li et al., 2024b). Recent years have witnessed the remarkable progress of large language models(Touvron et al., 2023; Chiang et al., 2023). By scaling up data size and model size, these LLMs raise amazing emergent abilities(Yin et al., 2023). This makes it possible to unify various vision and language tasks into a single framework. A prevalent paradigm for vision-language models involves adapting a vision encoder to a pretrained LLM with varying levels of integration, as done by LLaVA(Liu et al., 2024) and BLIP-2(Li et al., 2023). Vision-language models pre-trained on extensive visual-textual data exhibit exceptional performance across various downstream vision-language tasks. Additionally, there have been multiple efforts to adapt vision-language models for applications within the field of medical healthcare or specifically for radiology applications. For example, Med-Flamingo(Moor et al., 2023) is the first multimodal few-shot learner adapted to the medical domain, which promises novel clinical applications such as rationale generation. (Tu et al., 2024) introduced Med-PaLM M, a single multitask, multimodal biomedical AI system that can perform medical image classification and radiology report generation with the same set of model weights. LLaVA-Med(Li et al., 2024a) is a novel curriculum learning method for adapting LLaVA(Liu et al., 2024) to the biomedical domain using their self-generated biomedical multi-modal instruction-following dataset.

5 Conclusion

In this paper, we propose a method for RRG that utilizes a Cross-modal Enhancement and Alignment Adapter (CmEAA) to connect a vision encoder with a frozen large language model. In contrast

to methods that utilize a simple linear layer as the modality mapper, CmEAA harnesses contextual information from diverse observations in the training data to facilitate cross-modal feature enhancement of visual features and employs a Neural Mutual Information Aligner to effectively align global medical visual representations with textual representations. Experimental results on the MIMIC-CXR and IU X-Ray datasets demonstrate that our approach achieves a state-of-the-art performance and generates radiology reports that are both coherent and precise. Example of reports generated by our model indicates that CmEAA effectively bridge the gap between visual features and textual features. Ablation studies further validate the effectiveness of the proposed CmEAA.

Acknowledgments

This work is supported by the National Science and Technology Major Project (No.2021ZD0111000), Henan Provincial Science and Technology Research Project (No.232102211033).

References

- Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. 2018. Mutual information neural estimation. In *International conference on machine learning*, pages 531–540. PMLR.
- Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2022. Cross-modal memory networks for radiology report generation. *arXiv preprint arXiv:2204.13258*.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*.
- Pengyu Cheng, Weituo Hao, Shuyang Dai, Jiachang Liu, Zhe Gan, and Lawrence Carin. 2020. Club: A contrastive log-ratio upper bound of mutual information. In *International conference on machine learning*, pages 1779–1788. PMLR.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April 2023), 2(3):6.
- Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.
- Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587.
- Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.
- Monroe D Donsker and SR Srinivasa Varadhan. 1983. Asymptotic evaluation of certain markov process expectations for large time. iv. *Communications on pure and applied mathematics*, 36(2):183–212.
- Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. 2021. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR.
- Stacy K Goergen, Felicity J Pool, Tari J Turner, Jane E Grimm, Mark N Appleyard, Carmel Crock, Michael C Fahey, Michael F Fay, Nicholas J Ferris, Susan M Liew, et al. 2013. Evidence-based guideline for the written radiology report: Methods, recommendations and implementation challenges. *Journal of medical imaging and radiation oncology*, 57(1):1–7.
- Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. 2023a. Recap: Towards precise radiology report generation via dynamic disease progression reasoning. *arXiv preprint arXiv:2310.13864*.
- Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023b. Organ: observation-guided radiology report generation via tree reasoning. *arXiv preprint arXiv:2306.06466*.
- Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19809–19818.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.
- 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. 2019. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.
- Philippe Lambin, Ralph TH Leijenaar, Timo M Deist, Jurgen Peerlings, Evelyn EC De Jong, Janita

- Van Timmeren, Sebastian Sanduleanu, Ruben THM Larue, Aniek JG Even, Arthur Jochems, et al. 2017. Radiomics: the bridge between medical imaging and personalized medicine. *Nature reviews Clinical oncology*, 14(12):749–762.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.
- Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024b. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Haotian Liu, Chunyu Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022.
- Gen Luo, Minglang Huang, Yiyi Zhou, Xiaoshuai Sun, Guannan Jiang, Zhiyu Wang, and Rongrong Ji. 2023. Towards efficient visual adaptation via structural reparameterization. *arXiv preprint arXiv:2302.08106*.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakkas, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Hongyu Shen, Mingtao Pei, Juncai Liu, and Zhaoxing Tian. 2024. Automatic radiology reports generation via memory alignment network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4776–4783.
- Omkar Chakradhar Thawakar, Abdelrahman M Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. 2024. Xraygpt: Chest radiographs summarization using large medical vision-language models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 440–448.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. 2024. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138.
- A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.
- Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033.
- Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. 2018. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 457–466. Springer.
- Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.
- Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. 2019. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 721–729. Springer.
- Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020. When radiology report generation meets knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12910–12917.

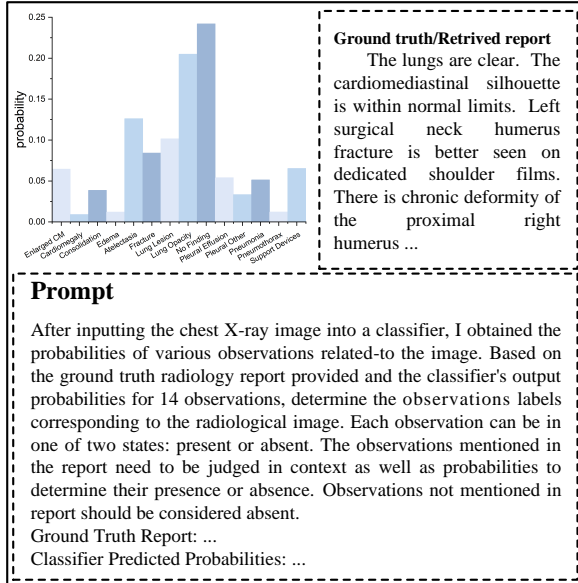


Figure 4: An example prompt we used to query the label of observation from GPT-4. The prompt includes the observation probabilities output by the classifier as well as the ground truth/retrieved report.

A Appendix

A.1 Observation Labels Extraction

There are 14 categories of observations: *No Finding*, *Enlarged Cardiomediastinum*, *Cardiomegaly*, *Lung Lesion*, *Lung Opacity*, *Edema*, *Consolidation*, *Pneumonia*, *Atelectasis*, *Pneumothorax*, *Pleural Effusion*, *Pleural Other*, *Fracture*, and *Support Devices*. Each observation label is classified as *Present*, *Absent*, or *Uncertain*. To simplify the prediction of observation labels, we regard *Present* and *Uncertain* as *Positive* and *Absent* as *Negative*. We employ cft-chexpert as the classifier to predict the probability of 14 observations corresponding to X-ray images. After obtaining the classification results, the probability of each observation, the prompt illustrated in Figure 4, and the ground truth report are input to GPT-4 to generate the observation labels for each image. For images within the validation and test sets, we use CLIP(Endo et al., 2021) pre-trained on MIMIC-CXR training set to retrieve similar reports from the training set as replacements for the ground truth reports.

A.2 Statistics of the Datasets

The statistics of two datasets are shown in Table 4, with the numbers of images, reports, and the average length of reports. The statistics of observation-related n-grams with different lengths extracted from two datasets are shown in Table 5. In IU X-

Dataset	IU X-Ray			MIMIC-CXR		
	Train	Val	Test	Train	Val	Test
Image #	5226	748	1496	368960	2991	5159
Report #	2770	395	790	222758	1808	3269
Avg.Len.	37.56	36.78	33.62	53.00	53.05	66.40

Table 4: The statistics of the two datasets, including the numbers of images, reports, and the average word-based length (Avg.Len.) of reports.

n-gram	Dataset	
	IU X-Ray	MIMIC-CXR
1-gram	1183	720
2-gram	238	838
3-gram	64	493
4-gram	-	287

Table 5: The statistics of observation-related n-grams with different lengths extracted from two datasets.

Ray dataset, 1-grams and 2-grams are predominant, significantly surpassing the number of 3-grams.

A.3 Implementation Details

We leverage Llama2-7B⁴ as the frozen large language model and the base version of Swin Transformer⁵ as the Visual Encoder. Based on previous research(Wang et al., 2023), for an input chest X-ray image X and its corresponding report R , the detailed prompt inputted into Llama2 is as follows:

$$\begin{aligned} \text{Human} &: < \text{Img} > X < / \text{Img} >, X_P. \\ \text{Assistant} &: R < / s > . \end{aligned}$$

Here X_P is the instruction prompt specific to the RRG task. In this paper, X_P ="Generate a comprehensive and detailed radiology report for this chest X-ray image." For this prompt, before inputting it into Llama2 for computation, X will be replaced by visual representations \tilde{X}^* processed by Eqn.3. Other text is tokenized into word tokens using Llama's tokenizer. The training process was conducted on one NVIDIA L20 48GB GPU using mixed precision for 3 epochs for MIMIC-CXR and 10 epochs for IU-Xray. For MIMIC-CXR dataset, we employed a mini-batch size of 6, with a learning rate of 1e-5. For IU-Xray, we employed a mini-batch size of 8, with a learning rate of 1e-4.

⁴<https://huggingface.co/meta-llama/Llama-2-7b-chat-hf>

⁵<https://huggingface.co/microsoft/swin-base-patch4-window7-224>