KIA: Knowledge-Guided Implicit Vision-Language Alignment for Chest X-Ray Report Generation

Heng Yin, Shanlin Zhou, Pandong Wang, Zirui Wu, Yongtao Hao*

Department of Computer Science and Technology, Tongji University, China
{yinheng, zhoushanlin, 1952124, 2111130, haoyt}@tongji.edu.cn

Abstract

Report generation (RG) faces challenges in understanding complex medical images and establishing cross-modal semantic alignment in radiology image-report pairs. Previous methods often overlook fine-grained cross-modal interaction, leading to insufficient understanding of detailed information. Recently, various large multimodal models have been proposed for image-text tasks. However, such models still underperform on rare domain tasks like understanding complex medical images. To address these limitations, we develop a new framework of **K**nowledge-guided **I**mplicit vision-language **A**lignment for radiology report generation, named **KIA**. To better understand medical reports and images and build alignment between them, multi-task implicit alignment is creatively introduced, forming comprehensive understanding of medical images and reports. Additionally, to further meet medical refinement requirements, we design novel masking strategies guided by medical knowledge to enhance pathological observation and anatomical landmark understanding. Experiments on two benchmark datasets show our KIA outperforms previous state-of-the-art methods in report quality and clinical efficacy.

1 Introduction

In modern medicine, analyzing radioactive medical images and writing reports are critical for diagnosing patient diseases. However, doctors spend considerable time manually performing this complex and laborious task, which may delay treatment (Bruno et al., 2015). Therefore, automatic report generation (RG) has garnered widespread research attention (Jing et al., 2017; Zhang et al., 2020; Akhter et al., 2023).

Unlike general image-to-text (I2T) generation tasks, RG requires stronger semantic alignment between images and reports, crucial for precisely identifying complex diseases. Some works attempt to establish medical associations between images and reports through explicit alignment. For instance, AlignTransformer (You et al., 2021) aligns image region-level features with several fixed disease tags. RGRG (Tanida et al., 2023) connects individual sentences to visual anatomical structures at the regional level, using manual annotation data. As shown in Fig. 1(a), these methods suffer from the need for extensive manual annotation and the lack of comprehensive feature interaction. So it is necessary to establish global and local semantic alignment with reduced labor annotation costs.

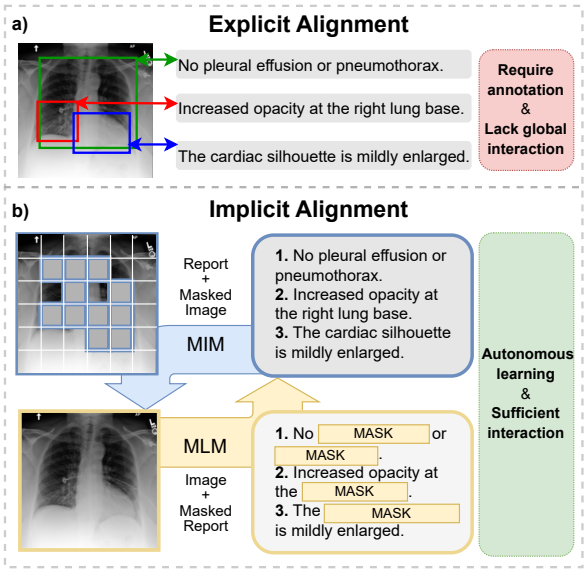

Figure 1: Explicit image-report alignment requires substantial densely annotated data and lacks global cross-modal interaction. In contrast, jointly masked multimodal modeling establishes implicit image-report alignment through autonomous learning.

Recently, several studies (Yan et al., 2021; Li et al., 2023a) employed contrastive learning to establish semantic associations through implicit alignment, which does not rely on fine-grained manually annotated data. Although they achieved some results, these associations remain limited to the

*Corresponding authors.

Proceedings of the 31st International Conference on Computational Linguistics, pages 4096–4108
January 19–24, 2025.

global level, proving insufficient for distinguishing complex diseases lacking obvious features. Now, large multimodal models (e.g., GPT-4V (Achiam et al., 2023), Gemini (Team et al., 2023)) have powerful image-to-text generation capabilities. However, these models exhibit low performance on complex tasks in rare domains such as medicine, which demands greater expertise.

In this paper, we propose a knowledge-guided implicit vision-language alignment framework for radiology report generation, called **KIA**. We achieve implicit alignment to promote the image-text bimodal interaction without relying on a large amount of additional manually annotated data. Specifically, KIA incorporates three high-level tasks: masked multimodal modeling (MMM), global alignment (GA) and report generation (RG), via joint multi-task training. To achieve local alignment, two low-level tasks are proposed within MMM: masked image modeling (MIM) and masked language modeling (MLM), as shown in Fig. 1b. These low-level tasks are inspired by joint masked tasks in vision-language pre-training (e.g., MaskVLM (Kwon et al., 2022), MAMO (Zhao et al., 2023)). To make full use of existing medical knowledge, MIM adopts a novel strategy of summarizing key patches for anatomical landmarks on the image and masking them to enhance learning of key features. MLM provides two new masking strategies, graph-guided and label-guided, that extract key information from image features. Leveraging global image-report information, we design GA, utilizing two low-level tasks, image-report contrastive learning (IRC) and image-report matching (IRM), for global modality alignment. Ultimately, RG is implemented with global and local information from above implicit alignment.

Our main contributions are as follows:

(1) We propose a novel framework, KIA, achieving comprehensive implicit alignment in image-report pairs via multi-task training without the need to manually annotate the data additionally.

(2) We propose a knowledge-guided masked multimodal modeling method, which summarizes knowledge of anatomical landmarks and disease observations of interest in healthcare and uses this knowledge to guide local implicit alignment.

(3) KIA achieves SOTA performance on MIMIC-CXR and IU-Xray datasets in both language generation and clinical efficacy metrics. We also discuss each alignment task in detail to demonstrate the effectiveness of bimodal interaction.

2 Related Work

2.1 Vision-Language Alignment

Vision-Language Alignment (VLA) can be divided into two categories: explicit and implicit alignment.

Explicit alignment methods (Li et al., 2020; Zeng et al., 2022) use densely labeled data to directly construct relationships between specific elements of the visual data and the corresponding parts of the textual data.

Implicit alignment methods enable models to discover local semantic associations between images and text on its own. Some methods (Radford et al., 2021; Li et al., 2021, 2022; Wang et al., 2022b; Huang et al., 2021) utilize contrastive learning for self-supervised image-text alignment, bringing positive pairs closer while pushing negative pairs apart. Some methods (Kwon et al., 2022; Singh et al., 2022; Zhao et al., 2023) have achieved masked multimodal modeling, promoting modal alignment through local random masking of images or text. With the emergence of multimodal large language models (MLLMs), some methods (Liu et al., 2024b; Bai et al., 2023) directly leverage generative pre-training tasks for alignment.

To the best of our knowledge, our method is the first to simultaneously apply multiple types of vision-language implicit alignment in the RG task.

2.2 Image Captioning and Radiology Report Generation

RG is the extension of image captioning (Vinyals et al., 2015; Anderson et al., 2018) in the medical domain. Many approaches attempt to adapt image captioning methods for RG. For instance, some methods (Jing et al., 2017; Yuan et al., 2019) generate medical reports leveraging visual feature extractors and language generators. Other methods enhance image and report understanding through feature alignment. AlignTransformer (You et al., 2021) aligns region-level image features with disease tags. RGRG (Tanida et al., 2023) establishes alignment between image regions and report phrases using object detection. ConVIRT (Zhang et al., 2022) and DCL (Li et al., 2023a) enhance global alignment via contrastive learning.

There are also some methods (Yang et al., 2022; Huang et al., 2023; Li et al., 2023a) that begin to use additional knowledge to assist in RG. Unlike these methods, we use this knowledge to guide implicit alignment.

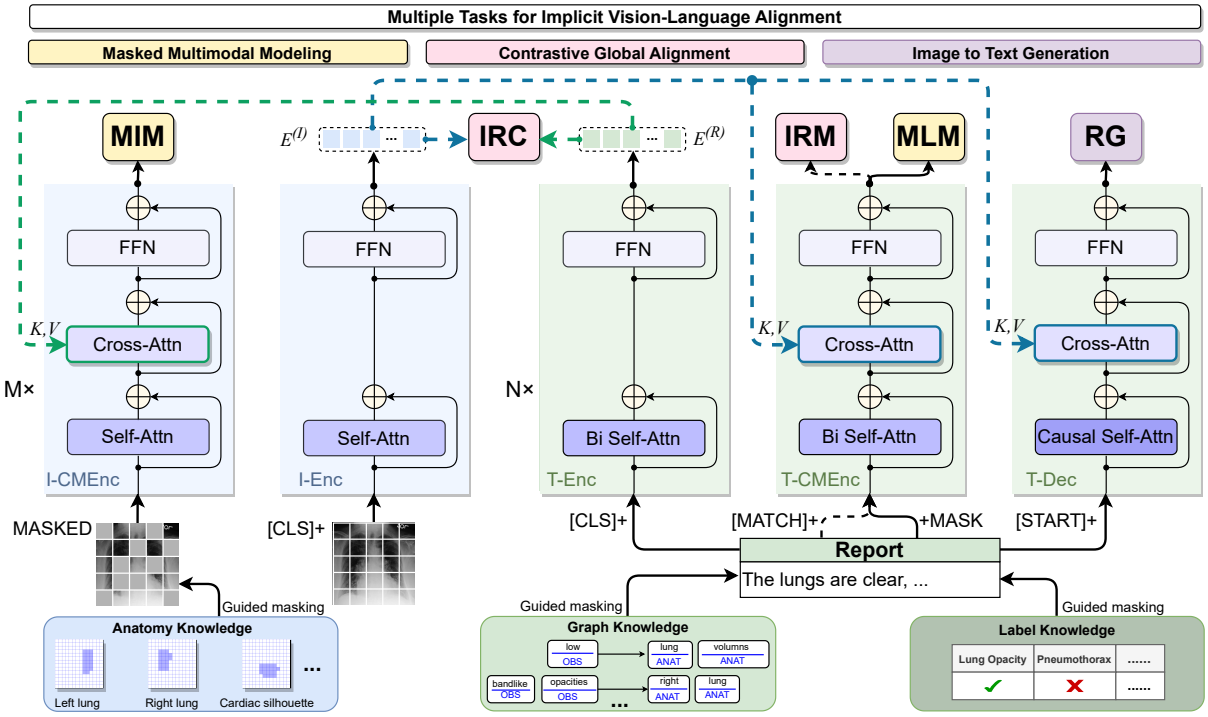

Figure 2: Illustration of our proposed KIA framework. Five components are used to perform unimodal and multimodal tasks: image (cross-modal) encoder (I-Enc and I-CMEnc), text (cross-modal) encoder (T-Enc and T-CMEnc) and text decoder (T-Dec). Five tasks (of three categories) are trained jointly in the framework, including masked image modeling (MIM), masked language modeling (MLM), image-report contrastive learning (IRC), image-report matching (IRM) and report generation (RG).

3 Methodology

In this section, we first present architecture in Sec. 3.1, and then introduce three implicit alignment tasks: knowledge-guided masked multimodal modeling in 3.2, contrastive global alignment in Sec. 3.3 and image to report generation in Sec. 3.4.

3.1 Model Architecture

Our architecture comprises five components for unimodal and multimodal tasks, as shown in Fig. 2.

Image Encoder. We utilize the Vision Transformer (ViT) (Dosovitskiy et al., 2020) as the image encoder. It divides an image I into patches and encodes them into a sequence of embeddings, incorporating a `[CLS]` token for global features.

Text Encoder. We use BERT (Devlin et al., 2018) as the text encoder. It encodes a sequence of text tokens of report R, incorporating a `[CLS]` token for global features.

Image (Text) Cross-Modal Encoder. Benefiting from the Transformer based architecture, the image (text) encoders can be easily extended to cross modal encoders by integrating cross-attention layers. Specifically, we enable bimodal interaction with an additional cross-attention layer between self-attention and FFN.

Text Decoder. The text decoder has the same structure as the text cross-modal encoder, except for replacing bi-directional attention with causal attention.

Dynamic Cross-Attention and Parameter Sharing. To enhance training efficiency, we reuse components across different modules as much as possible. Specifically, a cross attention layer is dynamically added to the unimodal encoder to perform multimodal tasks. A layer of computing operation in a component is:

$$X_l = \mathrm{LN}(\mathrm{SA}(X_{l-1}) + X_{l-1}) \tag{1a}$$

$$X_l = \begin{cases} \mathrm{LN}(\mathrm{CA}(\mathrm{X}_l, E^{(*)}) + X_l), & \text{if multimodal} \\ X_l, & \text{otherwise} \end{cases} \tag{1b}$$

$$X_l = \mathrm{LN}(\mathrm{FFN}(X_l) + X_l) \tag{1c}$$

where SA, CA, LN, and FFN respectively refer to the multi-head self-attention, multi-head cross-attention, layer normalization, and feed forward network modules(Vaswani et al., 2017). $E^{(*)}$ is the encoded feature sequence from the other modality.

3.2 Knowledge-Guided Masked Modeling

Instead of the random masking strategy used in previous methods, we use medical information, such as pathological observations and anatomical landmarks, as a guide for MMM.

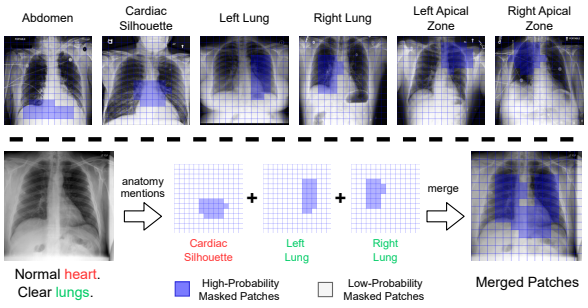

Figure 3: Given an image, we merge the key patches of mentioned landmarks in the report to obtain the final key masked patches.

Anatomy knowledge guided MIM (Anatomy-MIM). MIM enhances the image encoder's ability to extract features through self-supervised learning. However, existing unimodal MIM is unsuitable for medical images, since the pathology of the masked portion cannot be inferred solely from the unmasked portion. Thus, we utilize the image cross-modal encoder to interact with the report features to assist image reconstruction, thereby implicitly promoting semantic alignment between images and reports. Benefiting from the basically fixed position of the chest in radiological images, it is feasible to summarize key patches corresponding to anatomical landmarks, such as left lung and cardiac silhouette. We summarize the top 20 patch positions for 34 regions using a small number of images. Details of key patches are provided in the Appendix A. Then anatomy knowledge is used as a guide to mask the images. Masking key anatomy patches can compel the image cross-modal encoder to reconstruct the image based on report features, prompting the model to seek correspondences between image disease regions and text disease descriptions. Specifically, as shown in Fig. 3, for each image, we extract its anatomical landmarks mentioned in the report using simple word matching, then merge all key patches corresponding to these landmarks. The masking probabilities for different patches are as follows:

$$p_i = \begin{cases} p_{\text{key}}, & \text{if patch } i \text{ is in key anatomical locations} \\ p_{\text{other}}, & \text{otherwise} \end{cases} \quad (2)$$

$$N_{\text{key}} \cdot p_{\text{key}} + (N - N_{\text{key}}) \cdot p_{\text{other}} = N \cdot p_{\text{total}} \quad (3)$$

where N and N_{key} represent the total number of patches and the number of key patches, respectively. For p_{total}, we adopt the optimal experimental value of 0.6 from SimMIM (Xie et al., 2022), while p_{key} is elevated to a higher value, such as 0.9, to encourage the model to reconstruct the regions described in the report. In the experimental section we also present the impact of different p_{key} values on model performance. The MIM objective is defined as:

$$\mathcal{L}_{\text{MIM}} = \mathbb{E}_{(I,R)\sim D} \| I^M - f^M_{I-CMEnc}(I_m, \boldsymbol{\omega}) \|_1, \quad (4)$$

where $f_{I-CMEnc}(\cdot)$ is the image cross-modal encoder, I_m is the masked image, ω is the report features, and superscript M denotes that we only calculate the loss for the masked positions.

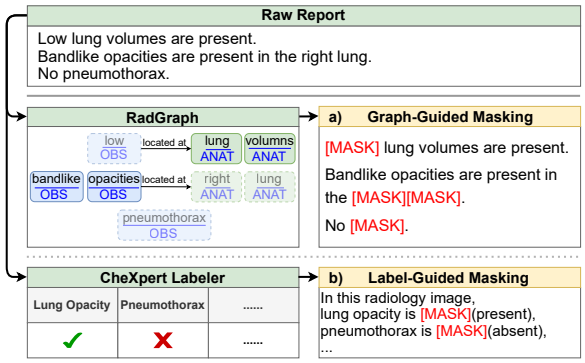

Figure 4: Graph-guided and label-guided masking strategies. The light-colored nodes of the graph represent the tokens to be masked. OBS and ANAT are abbreviations for observation and anatomy, respectively.

Graph knowledge guided MLM (Graph-MLM). The masked language modeling (MLM) task introduced by BERT (Devlin et al., 2018) has been widely used as a self-supervised pre-training task for natural language understanding. For general random text token masking methods, some masked tokens can be easily inferred based on the textual context. For example, in the sentence *"the heart size* `[MASK]` *within normal limits"*, the `[MASK]` token can easily be deduced as *"is"* solely based on the textual context. This simplicity does not facilitate encouraging the model to understand the connection between medical reports and images. So we design graph knowledge-guided MLM, as shown in Fig. 4a. In medical reports, most sentences take the form of describing the presence or absence of a disease symptom in a location in the image. We utilize RadGraph (Jain et al., 2021) tool to extract the raw

report into graph form as follows:[Observation] *located at* [Anatomy]. We use this graph as a knowledge guide to mask [Observation] part or [Anatomy] part of each sentence in the report randomly. This strategy forces the model to determine which [Anatomy] an [Observation] is located at or what [Observation] an [Anatomy] has, enhancing the interaction between images and reports.

Label knowledge guided MLM (Label-MLM). Some existing methods (Irvin et al., 2019; Smit et al., 2020) can extract disease labels from reports. We use disease label as a kind of guiding knowledge for multimodal alignment, i.e., label knowledge-guided MLM. Specifically, the report is labeled using the CheXpert (Irvin et al., 2019) tool, containing 14 common clinical abnormalities, e.g., atelectasis, lung opacity. Then we use these 14 labels to build a label report *"In this radiology image, atelectasis is present, cardiomegaly is absent, ... , lung opacity is absent."*. Finally we replace *present* and *absent* in the label report with the [MASK] token to obtain the masked report. This masking strategy forces the text model to query image information to determine whether an abnormality is *present* or *absent*.

For Graph-MLM and Label-MLM, the masked report is used as input to the text cross-modal encoder, interacting with image features for report reconstruction. The MLM objective is defined as:

$$\mathcal{L}_{\text{MLM}} = \mathbb{E}_{(I,R)\sim D}\mathcal{L}_{CE}(y_R^M, f_{T-CMEnc}^M(\boldsymbol{v}, R_m)) \quad (5)$$

where $f_{T-CMEnc}(\cdot)$ is the text cross-modal encoder, $\boldsymbol{v}$ is the image features, R_m is the masked report, y_R is the one-hot label from ground-truth text tokenization, $\mathcal{L}_{CE}(\cdot)$ is the cross-entropy loss function, and the superscript M denotes calculation of loss only at masked positions.

3.3 Contrastive Global Alignment

We also utilize two contrastive global alignment tasks to further promote modal alignment.

Image-Report Contrastive Learning (IRC). We follow ALBEF (Li et al., 2021) and BLIP (Li et al., 2022) using momentum contrastive learning approach. Similarities between an image I and report R, $s(I,R)$ and $s(R,I)$, are obtained after mapping them to the same dimensional space with linear projection. After softmax activation, we calculate image-to-report and report-to-image similarity: $p_k^{\text{I2R}}(I) = \frac{\exp(s(I,R_k)/\tau)}{\sum_{k=1}^{K}\exp(s(I,R_k)/\tau)}$ and $p_k^{\text{R2I}}(R)$, where τ is a learnable parameter and K represents the length of the queue in momentum-based contrastive learning. The IRC objective is defined as:

$$\begin{aligned}\mathcal{L}_{\text{IRC}} = \frac{1}{2}\mathbb{E}_{(I,R)\sim D}\big[&\mathcal{L}_{CE}(\boldsymbol{y}^{\text{I2R}}(I), \boldsymbol{p}^{\text{I2R}}(I)) \\ &+ \mathcal{L}_{CE}(\boldsymbol{y}^{\text{R2I}}(R), \boldsymbol{p}^{\text{R2I}}(R))\big]\end{aligned} \quad (6)$$

where $\mathcal{L}_{CE}$ is the cross-entropy loss function and $\boldsymbol{y}^*(\cdot)$ is ground truth of image-report similarity in one-hot form, based on constructed positive and negative pairs.

Image-Report Matching (IRM) predicts whether an image-report pair matches or not. In our architecture, both image and text cross-modal encoders can accomplish this task. For simplicity, we just use the text cross-modal encoder. We prefix the report sequence with a special token [MATCH] and feed it to the text cross-modal encoder. The embedding of the [MATCH] token is connected to a binary classification header to calculate the loss. The IRM objective is defined as:

$$\mathcal{L}_{\text{IRM}} = \mathbb{E}_{(I,R)\sim D}\mathcal{L}_{BCE}(\boldsymbol{y}^{\text{IRM}}, \boldsymbol{p}^{\text{IRM}}(I,R)) \quad (7)$$

where $\mathcal{L}_{BCE}$ is the binary cross-entropy loss function and $\boldsymbol{y}^{IRM}$ is a 2-dimensional one-hot form of ground-truth label.

3.4 Image to Report Generation

Image-to-text generation is a commonly used implicit alignment method in MLLMs (Bai et al., 2023; Liu et al., 2024b). Additionally, it is the ultimate task for the automatic medical report generation (RG). After prefixing a special token [START] to the report sequence, we feed it into the text decoder for language model autoregressive training. Report generation objective is defined as:

$$\mathcal{L}_{\text{RG}} = -\sum_{i=1}^{N}\log P(x_i|x_{<i}, \boldsymbol{v}) \quad (8)$$

where $\boldsymbol{v}$ represents the image features, and N is the number of report tokens.

The five tasks mentioned above are trained jointly in our framework. The final training objective is the weighted sum of the objectives of the five training tasks:

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{MIM}} + \lambda_2\mathcal{L}_{\text{MLM}} + \lambda_3\mathcal{L}_{\text{IRC}} + \lambda_4\mathcal{L}_{\text{IRM}} + \lambda_5\mathcal{L}_{\text{RG}} \quad (9)$$

where $\lambda_{\{1-5\}}$ are hyperparameters (all set to 1 by default).

Dataset	Methods	B-1	B-2	B-3	B-4	ROUGE-L	METEOR	CIDEr
IU-Xray	R2Gen (Chen et al., 2020)	0.470	0.304	0.219	0.165	0.371	0.187	-
	PPKED (Liu et al., 2021)	0.483	0.315	0.224	0.168	0.376	-	0.351
	R2GenCMN (Chen et al., 2022)	0.475	0.309	0.222	0.170	0.375	0.191	-
	MSAT (Wang et al., 2022a)	0.481	0.316	0.226	0.171	0.372	0.190	0.394
	DCL (Li et al., 2023a)	-	-	-	0.163	0.383	0.193	0.586
	METransformer (Wang et al., 2023)	0.483	0.322	0.228	0.172	0.380	0.192	0.435
	PromptMRG (Jin et al., 2024)	0.401	-	-	0.098	0.281	0.160	-
	BootstrapLLM (Liu et al., 2024a)	0.499	0.323	0.238	0.184	**0.390**	**0.208**	-
	KIA (Label)	0.501	0.325	0.240	0.183	0.375	0.207	0.559
	KIA (Graph)	**0.503**	**0.329**	**0.242**	**0.188**	0.385	**0.208**	**0.641**
MIMIC-CXR	R2Gen (Chen et al., 2020)	0.353	0.218	0.145	0.103	0.277	0.142	-
	PPKED (Liu et al., 2021)	0.360	0.224	0.149	0.106	0.284	0.149	-
	R2GenCMN (Chen et al., 2022)	0.353	0.218	0.148	0.106	0.278	0.142	-
	MSAT (Wang et al., 2022a)	0.373	0.235	0.162	0.120	0.282	0.143	0.299
	KiUT (Huang et al., 2023)	0.393	0.243	0.159	0.113	0.285	0.160	-
	UAR (Li et al., 2023b)	0.363	0.229	0.158	0.107	0.289	0.157	0.289
	PromptMRG (Jin et al., 2024)	0.398	-	-	0.112	0.268	0.157	-
	BootstrapLLM (Liu et al., 2024a)	0.402	0.262	0.180	0.128	0.291	**0.175**	-
	KIA (Label)	0.413	0.272	0.185	0.136	0.305	0.164	0.307
	KIA (Graph)	**0.415**	**0.274**	**0.187**	**0.138**	**0.307**	0.167	**0.316**

Table 1: Comparison with state-of-the-art methods on IU X-Ray and MIMIC-CXR datasets. Label represents using label-guided MLM and Graph represents using graph-guided MLM. The best results are in **boldface** and the second best results are underlined. B-{1-4} are abbreviations for BLEU-{1-4}.

Methods	Precision	Recall	F1-Score
R2Gen	0.333	0.273	0.276
KnowMat	0.458	0.348	0.371
KiUT	0.371	0.318	0.321
DCL	0.471	0.352	0.373
KIA w/o MIM	0.496	0.392	0.438
KIA w/o MLM	0.486	0.359	0.413
KIA (ours)	**0.504**	**0.425**	**0.461**

Table 2: Comparison of CE metrics on MIMIC-CXR dataset. w/o is the abbreviation of without.

4 Experiments

4.1 Datasets

Experiments are conducted on two widely used datasets: IU-Xray and MIMIC-CXR.

IU-Xray (Demner-Fushman et al., 2016) contains 7,470 radiology images and 3,955 reports. Each report corresponds to either a frontal view or a combination of a frontal and a lateral view. We adopt the same data partitioning as Chen et al. (2020) for a fair comparison.

MIMIC-CXR (Johnson et al., 2019) is the largest radiology dataset, containing 377,110 images and 227,835 reports. We use the official train/val/test splits following Chen et al. (2020).

4.2 Evaluation Metrics and Settings

We apply the most widely used natural language generation (NLG) metrics and clinical efficacy (CE) metrics to evaluate our model's performance.

NLG Metrics include BLEU-n (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Chin-Yew, 2004), and CIDEr (Vedantam et al., 2015). We use the MS-COCO caption evaluation tool [1] to calculate these metrics.

CE Metrics , recently proposed to measure the clinical correctness of generated reports, utilize the CheXpert (Irvin et al., 2019) labeling tool to annotate 14 clinical abnormality categories in the generated reports. We calculate precision, recall, and F1 scores using the ground truth. As the IU-XRay dataset lacks officially available CheXpert labeled data, we only calculate CE metrics on the MIMIC-CXR dataset.

Implementation Details. We use ViT/B16 (Dosovitskiy et al., 2020) as the image encoder and a pre-trained BERT (Devlin et al., 2018) as the text encoder. The training process is divided into two stages. In stage one, all five tasks are trained jointly for 40/10 epochs on the IU-Xray/MIMIC-CXR dataset. In stage two, only the report generation task is further trained for an additional 5 epochs to adapt to the final task. The model is trained on a single NVIDIA 4090 GPU with a batch size of 32. The learning rate is initialized at 1e-4 and undergoes a linear decay with a decay rate of 0.98. The optimizer

[1] https://github.com/tylin/coco-caption

is RAdam (Liu et al., 2019) with a weight decay of 0.05. All hyperparameters are tuned on the validation set, using the checkpoint with the best CIDEr score for testing.

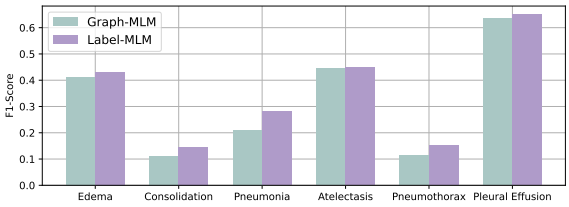

Figure 5: The clinical efficacy F1-score for six clinical abnormalities using {Graph,Label}-MLM.

4.3 Main Results

Baselines. We conduct a quantitative comparison using existing SOTA methods. These methods can be categorized into four groups: memory-driven models, including R2Gen (Chen et al., 2020) and R2GenCMN (Chen et al., 2022); medical/expert knowledge-enhanced models, including PPKED (Liu et al., 2021), MSAT (Wang et al., 2022a), DCL (Li et al., 2023a), KiUT (Huang et al., 2023), and METransformer (Wang et al., 2023); cross-modal alignment enhancement models, including UAR (Li et al., 2023b); and large language model-based medical models, including BootstrapLLM (Liu et al., 2024a). Additionally, we perform a qualitative comparison with some general multimodal large language models, including GPT-4V (Achiam et al., 2023).

NLG Results. The results in Tab. 1 show that our KIA outperforms existing methods on most NLG metrics and achieves SOTA performance. Specifically, our KIA achieves 0.188 and 0.138 BLEU-4 on IU-XRay and MIMIC-CXR, respectively (2% and 8% improvement). In terms of ROUGE-L and METEOR, our method also shows competitive performance. Notably, our CIDEr scores of 0.641 and 0.316 on two datasets achieve considerable improvements of 9% and 6%, showing that reports generated by KIA are highly relevant to the image content and accurately reflect important information. This illustrates that the multi-task implicit alignment we designed can effectively enhance the understanding of images and reports.

Clinical Correctness. We also evaluate CE metrics on MIMIC-CXR dataset of our method in comparison to other baseline methods. As shown in Tab. 2, our method significantly outperforms previous state-of-the-art methods. The improvement is attributed to our strategy of using graph-guided, label-guided and anatomy-guided masking, enabling the model to fully learn crucial medical information. It is worth noting that, although the label-guided masking strategy performs lower than the graph-guided masking strategy on NLG metrics, it exhibits better performance on CE metrics as shown in Fig. 5. This demonstrates that performing MLM based on clinical abnormalities can enhance clinical correctness.

4.4 Ablation Study

In this section, we conduct ablation studies to investigate the contribution of each task in our proposed KIA. Tab. 3 shows the quantitative analysis of NLG metrics for different combinations of tasks.

Effect of Anatomy-MIM. The Anatomy-MIM (setting *b*) leads to an improvement in NLG metrics over general MIM (setting *a*), with an average increase of +6.5% compared to +5.9%. In addition, we conduct experiments under multiple MIM mask probability combinations. In the Appendix B we demonstrate the performance of the model with different combinations of masking probabilities, which exemplifies the effectiveness of masking key regions with high probability.

Effect of {Graph, Label}-MLM. We evaluate three distinct masking strategies: general MLM, Graph-MLM and Label-MLM. The Graph-MLM (setting *d*) consistently outperforms general MLM (setting *c*) and closely rivals Label-MLM (setting *e*). The superior performance of Graph-MLM, with an average NLG metric increase of +6.3%, underlines the ability of capturing the key medical information from the image. Tab. 2 shows that the model has some decrease in CE metrics in the absence of the MLM, demonstrates the significant effect of using the MLM task to establish alignment between images and reports.

Effect of multi-task learning. The model settings *f*, *g* and the proposed KIA model illustrate the benefits of multi-task learning. The culmination of the multi-task learning approach is best exemplified in the KIA (our proposed), which includes all tasks with optimized masking strategies. It achieves the highest increase in performance metrics, with an average NLG improvement of +15.1%.

Effect of parameter sharing. The Tab. 4 indicates that parameter sharing can reduce the number of model parameters from 283M to 120M, enhancing the training efficiency. The regularization effect

Settings	MIM	MLM	IRC	IRM	RG	BLEU-4	ROUGE-L	METEOR	CIDEr	AVG.Δ
Base					✓	0.164	0.361	0.200	0.474	-
(a)	✓(General)				✓	0.171	0.373	0.201	0.547	+5.9%
(b)	✓(Anatomy)				✓	0.172	0.372	0.203	0.553	+6.5%
(c)		✓(General)			✓	0.168	0.361	0.206	0.525	+4.0%
(d)		✓(Graph)			✓	0.173	0.365	0.205	0.550	+6.3%
(e)		✓(Label)			✓	0.171	0.360	0.203	0.538	+4.7%
(f)	✓(Anatomy)	✓(Graph)			✓	0.179	0.374	0.207	0.588	+10.1%
(g)			✓	✓	✓	0.176	0.379	0.203	0.545	+7.2%
KIA	✓(Anatomy)	✓(Graph)	✓	✓	✓	**0.188**	**0.385**	**0.208**	**0.641**	+15.1%

Table 3: Ablation results on IU-Xray dataset using different combinations of tasks, where "General" stands for "General masking", "Anatomy" stands for "Anatomy-MIM", "Graph" stands for "Graph-MLM", and "Label" stands for "Label-MLM". The "AVG.Δ" column presents the average improvement of all NLG metrics.

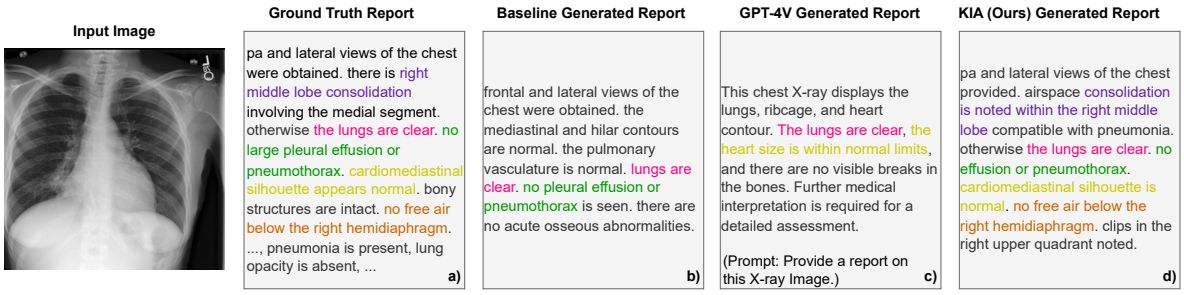

Figure 6: Examples of reports generated by the baseline method, GPT-4V and proposed KIA. Sentences corresponding to ground-truth report in the generated report are annotated with the same color.

Methods	#params	BLEU-4	CIDEr
w/o Parameter Sharing	283M	0.166	0.456
w/ Parameter Sharing	120M	**0.188**	**0.641**

Table 4: Comparison of model performance and parameter quantity with and without parameter sharing.

of parameter sharing improves the model's BLEU-4 metric on IU-Xray from 0.166 to 0.188.

Analysis of each component. Our KIA implements multiple image-text implicit alignment tasks within a simple model architecture. The MIM and MLM tasks primarily focus on aligning local features, which is crucial for medical images that emphasize specific local areas of disease. Moreover, the guidance of medical knowledge can further enhance this alignment effect, as shown in Tab. 3(a-e). Moreover, the final medical report generation requires a summary diagnosis from a global perspective, making global alignment tasks equally important for chest X-ray report generation, as indicated in Tab. 3(g). Ultimately, multi-task collaborative training simultaneously enhances the model's ability to extract both local and global features, thereby maximizing model performance.

4.5 Qualitative Analysis

We conduct qualitative analysis from various perspectives. First, we compare reports generated by different methods. Second, we also visualize cross-modal attention in masked modeling. To verify the effectiveness of implicit alignment of image and text, we conduct qualitative analysis of masked modeling and globa alignment. Due to space limitations, we put this part in the Appendix C.

Report Generation. Fig. 6(b-d) shows the reports generated by the baseline (only RG is trained), GPT-4V (Achiam et al., 2023) and our method. Comparison to ground-truth report reveals that our model extracts more medical diagnostic information from images than baseline model and GPT-4V.

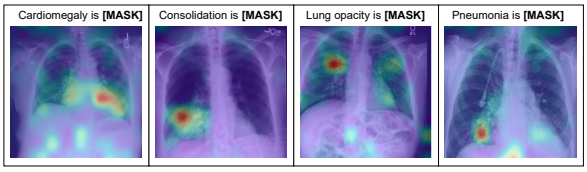

Figure 7: Attention visualization of [MASK] token on images. The ground-truth of [MASK] token is "present" or "absent".

Attention Visualization. To explore model interpretability, we visualize attention maps of [MASK] tokens interacting with images in label-guided MLM, as shown in Fig. 7. Highlighted regions indicate where the model places its attention when determining [MASK] token value, which is categorized as "present" or "absent" for clinical abnormalities like cardiomegaly, consolidation, etc. The attention heatmap demonstrates our model indeed focuses attention on key locations during MLM, showing effective alignment.

5 Conclusion

In this paper, we propose KIA, a method leveraging masked multimodal modeling, contrastive alignment and image-to-text tasks to facilitate comprehensive implicit alignment between radiology images and reports. By performing multi-task training in a simple framework, the model's understanding of medical images significantly improves, enhancing report generation. With our designed knowledge guided masked modeling, the image and report modalities are fully interacted, improving extraction of critical medical information. Experimental results on two benchmarks show our proposed framework effectively contributes to automatic radiology report generation quality.

6 Limitation

Our research has achieved image-text alignment in the chest X-ray field to generate chest X-ray reports. It should be noted that we have not yet conducted experiments in other medical domains beyond chest X-rays. Due to the scarcity of data in other medical fields, we plan to fine-tune the pre-trained KIA model with a small amount of data from other medical domains in future work to validate the generalization capability of our approach.

In addition, our method can train a better medical image feature encoder, but we did not introduce a larger text decoder. In the future, we will further expand our method on large language models with more than 7B parameters.

7 Ethical Consideration

Our study used two chest X-ray datasets, IU-Xray and MIMIC-CXR, which are publicly available for scientific research. They were downloaded from the official dataset website. We followed the ethical consideration of the source dataset and only used the dataset for scientific research in this paper, not for any other purpose.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yasmeena Akhter, Richa Singh, and Mayank Vatsa. 2023. Ai-based radiodiagnosis using chest x-rays: A review. *Frontiers in Big Data*, 6:1120989.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Satanjeev Banerjee and Alon Lavie. 2005. 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.

Michael A Bruno, Eric A Walker, and Hani H Abujudeh. 2015. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*, 35(6):1668–1676.

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*.

Lin Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*.

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.

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. *arXiv preprint arXiv:2020.11929*.

Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951.

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, Silviana 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.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2607–2615.

Baoyu Jing, Pengtao Xie, and Eric Xing. 2017. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*.

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*.

Gukyeong Kwon, Zhaowei Cai, Avinash Ravichandran, Erhan Bas, Rahul Bhotika, and Stefano Soatto. 2022. Masked vision and language modeling for multi-modal representation learning. *arXiv preprint arXiv:2208.02131*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023a. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

Yaowei Li, Bang Yang, Xuxin Cheng, Zhihong Zhu, Hongxiang Li, and Yuexian Zou. 2023b. Unify, align and refine: Multi-level semantic alignment for radiology report generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2863–2874.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024a. Bootstrapping large language models for radiology report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18635–18643.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13753–13762.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. 2019. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus

Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11558–11567.

Zhanyu Wang, Mingkang Tang, Lei Wang, Xiu Li, and Luping Zhou. 2022a. A medical semantic-assisted transformer for radiographic report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–664. Springer.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022b. Medclip: Contrastive learning from unpaired medical images and text. In *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*.

Zhenda Xie, Zheng Zhang, Yue Cao, Yutong Lin, Jianmin Bao, Zhuliang Yao, Qi Dai, and Han Hu. 2022. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9653–9663.

An Yan, Zexue He, Xing Lu, Jiang Du, Eric Chang, Amilcare Gentili, Julian McAuley, and Chun-Nan Hsu. 2021. Weakly supervised contrastive learning for chest x-ray report generation. *arXiv preprint arXiv:2109.12242*.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. 2022. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, pages 72–82. Springer.

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.

Yan Zeng, Xinsong Zhang, and Hang Li. 2022. Multi-grained vision language pre-training: Aligning texts with visual concepts. In *International Conference on Machine Learning*, pages 25994–26009. PMLR.

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.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR.

Zijia Zhao, Longteng Guo, Xingjian He, Shuai Shao, Zehuan Yuan, and Jing Liu. 2023. Mamo: Fine-grained vision-language representations learning with masked multimodal modeling. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1528–1538.

A Key Anatomical Patches

We summarize anatomical positional knowledge based on a small amount of annotated data, namely key patches corresponding to 34 anatomical positions, as shown in Fig. 8. Due to the positional

stability of chest X-ray images, these patches have a high degree of adaptability on different images. Naturally, we do not need these patches to completely and accurately cover a specific part of an image like object detection datasets, as they are only used as prior knowledge to guide MIM implicit alignment tasks.

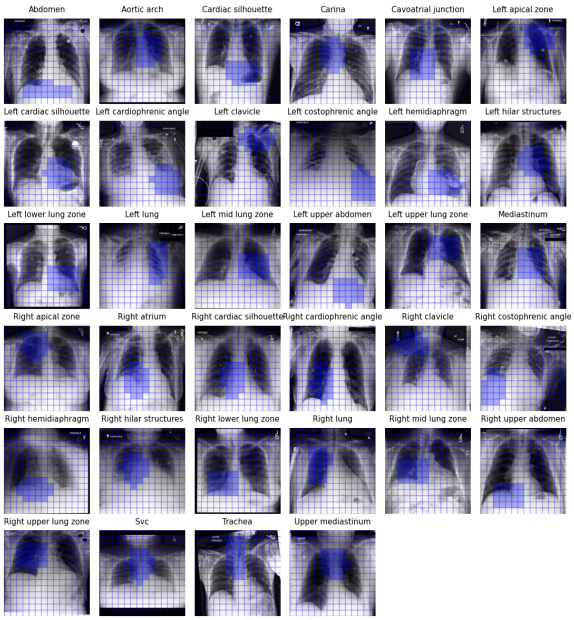

Figure 8: Key patches corresponding to 34 anatomical landmarks. Each landmark corresponds to 20 key patches. (Zoom to view)

B Different Image Mask Probabilities

Total mask prob	Key mask prob	BLEU-4	CIDEr
0.6	1.0	0.137	0.314
0.6	0.8	0.136	0.314
0.6	0.6	0.131	0.308
0.6	0.4	0.131	0.301
0.9	0.9	0.135	0.310
0.6	0.9	**0.138**	**0.316**

Table 5: Performance on the MIMIC-CXR dataset with different combinations of mask probabilities. The combination of 0.6+0.9 is the best experimental result we obtained.

As shown in Tab 5, the total mask probability of 0.6 and the key patch of 0.9 yield the best performance, improving by 2.5% on CIDEr compared to the general 0.6 mask probability. High-probability masking in key regions forces the model to obtain information from the report for reconstruction, while avoiding simple reconstruction based on the surrounding patches.

C More Qualitative Analysis

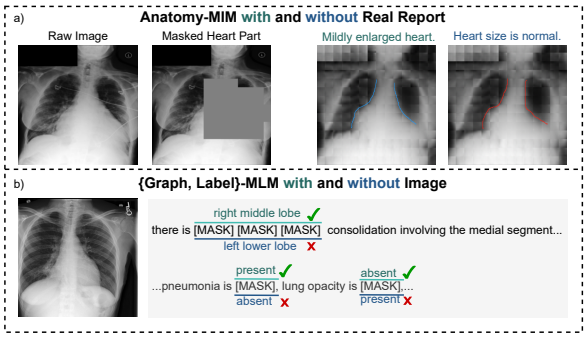

Figure 9: Example of cross-modal interaction in MMM.

Masked Modeling. To verify the effectiveness of masked modeling alignment, we construct experiments under different circumstances. Specifically, as shown in Fig. 9a, with different reports given, the reconstruction of the heart part of the image differs and corresponds to the semantics of the report, illustrating our MIM modeling interacts across modalities. The effects of MLM modeling with and without images shown in Fig. 9b similarly illustrate that the masked modeling of text is also based on image modality information. The cross-modal interaction is the source of the ability of the final report generation task to understand the semantics of the images.

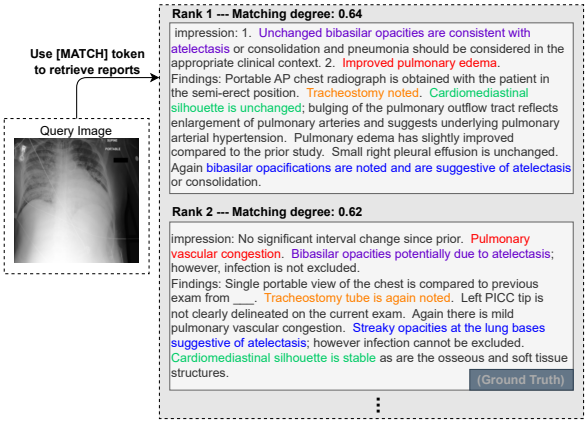

Figure 10: Example of image-text retrieval.

Global Alignment. We utilize the [MATCH] token in the IRM task for binary classification to judge whether an image-text pair matches. Based on the classification probability, we obtain a global matching degree for the image-text pair, and thereby implement a simple image-to-text retrieval task. As shown in Fig. 10, for a given image, although the ground truth is not the top in the retrieval result, the retrieved highly matched reports

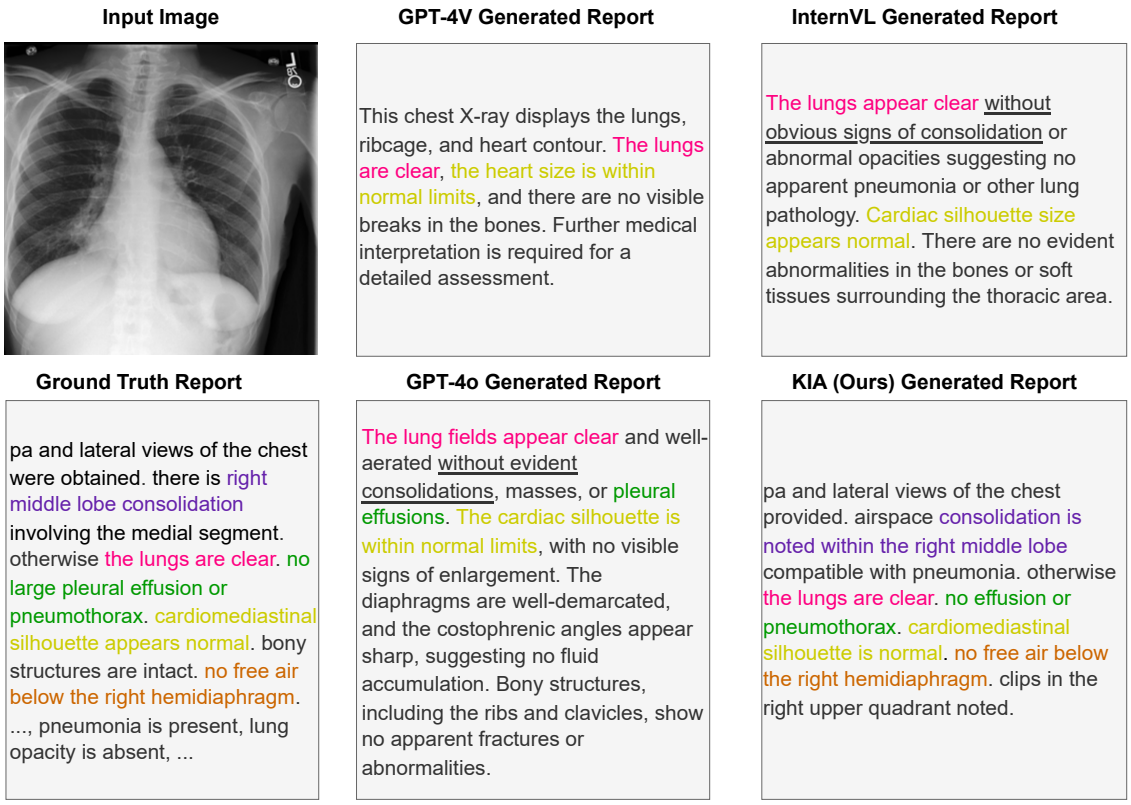

Figure 11: More qualitative results of different multimodal large language models including GPT-4V, GPT-4o and InternVL. Sentences corresponding to ground-truth report in the generated report are annotated with the same color. The underlined statement indicates inconsistency with the narrative in the ground truth report.

are indeed extremely similar, showing our method effectively aligns the global semantics of medical images and reports.

D Qualitative Results of more MLLMs

We test more existing advanced multimodal large language models, including GPT-4V, GPT-4o, and InternVL, as shown in Fig. 11. Although general multimodal large models can provide report-style descriptions of chest X-ray images, they struggle with handling certain disease details. For example, both GPT-4o and InternVL incorrectly conclude that there is no consolidation in the image, which contradicts the actual report. The reports generated by our KIA method align more closely with the content of the ground truth report.

E Error Case Analysis

We show an error case in Fig 12, revealing inconsistencies in report generated by KIA versus the ground truth (GT) report. We respectively calculate the matching degree of the GT report and the report generated by KIA with the [MATCH] token in the IRM task, and observe that the matching degree of

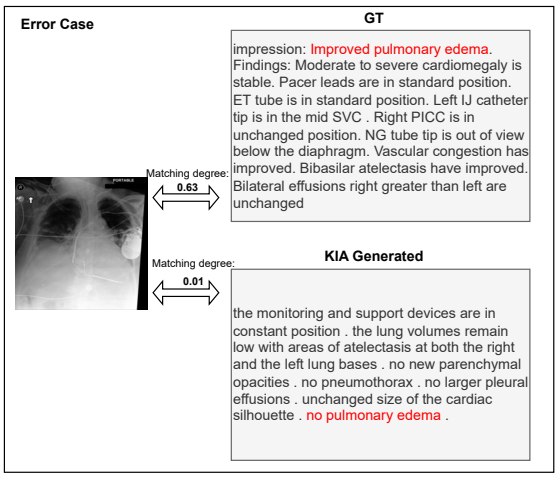

Figure 12: Example of error case.

GT is much higher than that of the report generated. This indicates that there is still a situation of poor generation when the alignment is good. We believe that this is because text generation requires a much larger model compared to the understanding task. Therefore, on the basis of our well-aligned model, introducing LLM as the text generator is a potential research.