Fine-Grained Image-Text Alignment in Medical Imaging Enables Explainable Cyclic Image-Report Generation

Wenting Chen[1]   **Linlin Shen**[3]   **Jingyang Lin**[4]   **Jiebo Luo**[4]
Xiang Li[5*]   **Yixuan Yuan**[2*]

[1]City University of Hong Kong   [2]The Chinese University of Hong Kong
[3]Shenzhen University   [4]University of Rochester
[5]Massachusetts General Hospital and Harvard Medical School

[1]wentichen7-c@my.cityu.edu.hk   [2]yxyuan@ee.cuhk.edu.hk   [3]llshen@szu.edu.cn
[4]{jluo@cs, jlin81@ur}.rochester.edu   [5]xli60@mgh.harvard.edu

Abstract

Fine-grained vision-language models (VLM) have been widely used for inter-modality local alignment between the predefined fixed patches and textual words. However, in medical analysis, lesions exhibit varying sizes and positions, and using fixed patches may cause incomplete representations of lesions. Moreover, these methods provide explainability by using heatmaps to show the general image areas potentially associated with texts rather than specific regions, making their explanations not explicit and specific enough. To address these issues, we propose a novel Adaptive patch-word Matching (AdaMatch) model to correlate chest X-ray (CXR) image regions with words in medical reports and apply it to CXR-report generation to provide explainability for the generation process. AdaMatch exploits the fine-grained relation between adaptive patches and words to provide explanations of specific image regions with corresponding words. To capture the abnormal regions of varying sizes and positions, we introduce an Adaptive Patch extraction (AdaPatch) module to acquire adaptive patches for these regions adaptively. Aiming to provide explicit explainability for the CXR-report generation task, we propose an AdaMatch-based bidirectional LLM for Cyclic CXR-report generation (AdaMatch-Cyclic). It employs AdaMatch to obtain the keywords for CXR images and 'keypatches' for medical reports as hints to guide CXR-report generation. Extensive experiments on two publicly available CXR datasets validate the effectiveness of our method and its superior performance over existing methods.

1   Introduction

Inter-modality alignment, such as vision and language, has been an important task with growing interests in the field of computer vision, especially with the recent advancement in representation
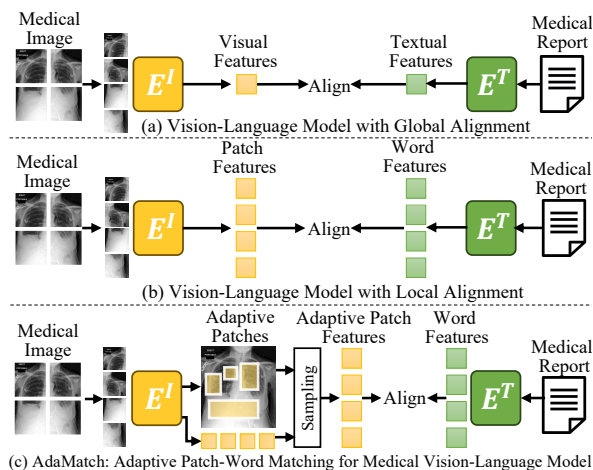

Figure 1: Current vision-language models (VLM) achieve (a) global alignment and (b) local alignment by matching overall visual with textual features, and aligning patches with word features, respectively. (c) To exploit the relation between textual words and abnormal patches with varied sizes, our AdaMatch obtains adaptive patch features and aligns them with word features.

learning (Radford et al., 2021). Technologies like contrastive learning and self-supervised learning have dramatically improved state-of-the-art alignment performance. Recent vision-language models (VLMs) demonstrate two approaches: global contrastive alignment, which integrates images and texts at a global level (Radford et al., 2021; Jia et al., 2021; Jang et al., 2023; Wang et al., 2023; Yang et al., 2022), and local alignment, focusing on detailed connections between visual objects and textual words (Chen et al., 2020a; Li et al., 2020b,a; Zhan et al., 2021; Kim et al., 2021; Yao et al., 2021), as illustrated in Fig. 1.

Current VLMs with local alignment either adopt the pre-trained object detector to extract region-of-interest (ROI) features from images and match the corresponding object features with textual words (Chen et al., 2020a; Li et al., 2020b,a; Zhan et al., 2021), or align the visual token from each

---

*Xiang Li and Yixuan Yuan are corresponding authors.

patch and the textual token into the same embedding space (Kim et al., 2021; Yao et al., 2021; Ji et al., 2021; Wang et al., 2022a). The former highly relies on the quality of the object detector and its predefined classes, which is less generalizable to new domains. The latter family of methods learns the alignment in a more automatic and data-driven manner. However, most of these methods depend on a pre-defined patch size and positions (e.g., grids) across images. In the most challenging cases, such as the analysis of medical image, lesions can exhibit a wide range of shapes, sizes, and positions. A fixed partition of image patches can lead to incomplete or ambiguous representations of the key imaging abnormalities. Therefore, it is highly desirable to adaptively exploit the fine-grained relationship between image embeddings derived from a more flexible patching scheme and textual embeddings.

Another challenge in the current VLMs lies in their explainability: it is generally difficult to delineate the image-text relationship learned by the model, especially for the current medical VLMs. Current solutions to provide such explanations in medical VLMs leverage the attention maps from the intermediate layer to visualize the location of the abnormalities (Moon et al., 2022; Huang et al., 2021; Yan and Pei, 2022). Other methods (Wan et al., 2023; Chen et al., 2023; Liu et al., 2023) utilize network gradients such as Grad-CAM (Selvaraju et al., 2017) to generate the heatmaps according to lesion types based on ground-truth reports. However, both maps can only show the general areas potentially associated with the corresponding text data rather than pinpointing a specific region. In addition, gradient-based methods need ground-truth reports, prohibiting them from functioning correctly beyond training data. It is thus highly necessary to develop a mechanism that could provide explicit and specific explanations of input image or text during inference time.

To address these two challenges above, we propose a novel Adaptive patch-word Matching (AdaMatch) model to match fine-grained image regions of various sizes and positions with textual data. AdaMatch introduces an image encoder with multiple Adaptive Patch extraction (AdaPatch) modules to adaptively acquire the patches associated with certain text tokens. It then performs patch-word alignment based on contrastive learning. AdaMatch is specifically developed in the context of aligning radiology images (chest X-ray, CXR) and their corresponding radiology reports with the capability of achieving cyclic (CXR-to-report and report-to-CXR) generation based on the learned alignment. Our premise is that such a cyclic generation task would serve as the best use case and evaluation criterion for the desired fine-grained alignment. Also, fine-grained cyclic generation between CXR and report will provide natural explainability for how the model aligns two modalities: for any given text token, we can visualize its matching imaging manifestation; and for any image region within a CXR image, we can tell the type of lesion or anatomical region it belongs to.

To implement the cyclic CXR-report generation, we propose an AdaMatch-based bidirectional model (AdaMatch-Cyclic). AdaMatch-Cyclic employs AdaMatch to identify the keywords for CXR images and the 'keypatches' for medical reports to guide the generation tasks. Since the potential keywords for CXR images cover a wide range and ground-truth reports cannot be used during inference, we predefine a textual codebook with the most common entities from medical reports as prior knowledge during fine-grained alignment. With the textual codebook, AdaMatch aligns it with the adaptive patches to obtain matched keywords to facilitate report generation. Next, a VQ-GAN model encodes the CXR image into image tokens, and a Large Language Model (LLM) takes image tokens, the matched keywords, and the instructions as input to generate medical reports. Similarly, we also build a visual codebook with the most commonly seen patches as 'keypatches', and use AdaMatch to obtain the matched keypatches from given text reports as hints for CXR generation. Utilizing medical reports, matched keypatches, and instructions, LLM generates image tokens, subsequently decoded by the VQ-GAN model to produce the resulting CXR image. Our contributions are summarized as follows:

- To exploit the fine-grained relation between CXR image patches and words of medical reports, we propose an Adaptive patch-word Matching (AdaMatch) model to obtain adaptive patches for abnormal regions and perform alignment between them and texts in medical reports.

- We devise an AdaMatch-based bidirectional LLM for Cyclic CXR-report generation (AdaMatch-Cyclic) to facilitate the bidirectional generation between CXR and re-
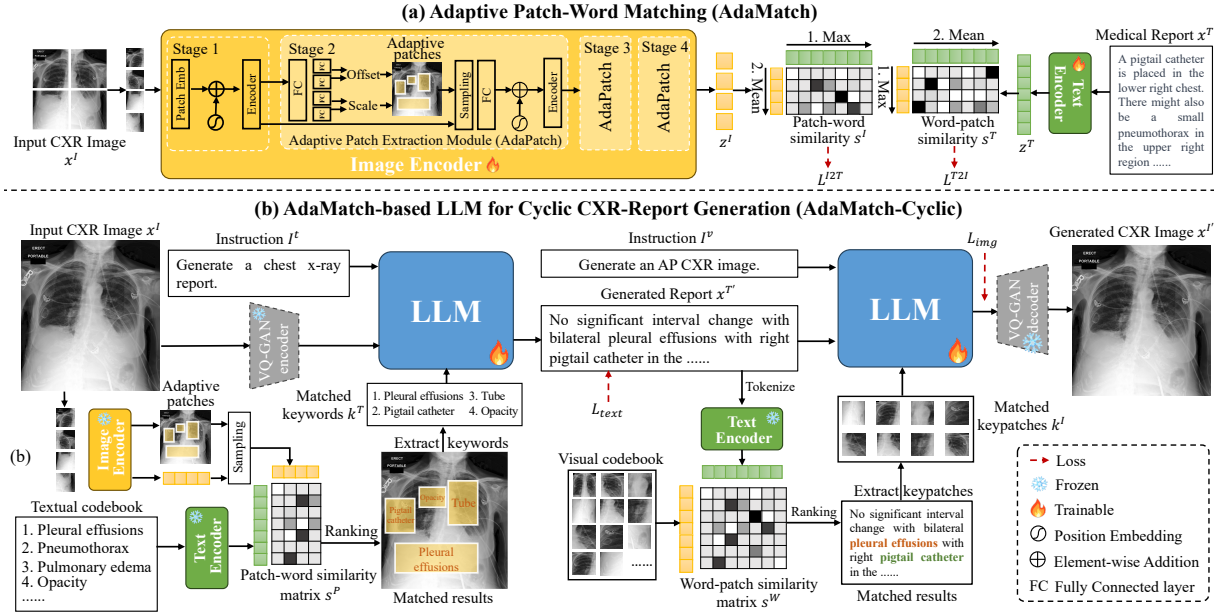
Figure 2: The overview of the proposed methods. (a) Adaptive patch-word Matching (AdaMatch) model. (b) AdaMatch-based bidirectional large language model (LLM) for cyclic CXR-report generation (AdaMatch-Cyclic).

ports. Moreover, we build the textual and visual codebook to utilize AdaMatch to extract useful keywords and keypatches for the report and CXR generation, respectively.

- Experiments on two publicly available chest X-ray datasets demonstrate the effectiveness of our method and its superior performance over the state-of-art methods.

2 Related Works

2.1 Fine-Grained Vision-Language Models

Recently, several fine-grained vision-language models (VLM) achieve the local alignment by exploiting the fine-grained relation between visual objects and textual words. Some methods (Chen et al., 2020a; Li et al., 2020b,a; Zhan et al., 2021) employ the pre-trained object detector to obtain object features from images and align them with textual features, and others (Kim et al., 2021; Yao et al., 2021; Wang et al., 2022a; Ji et al., 2021; Xu et al., 2023; Jiang et al., 2023) aim to align the fixed patches with the textual words locally. The former relies on a precise object detector, and the latter focuses on the relation between fixed patches inside the predefined grid and words. However, varying lung lesion characteristics may cause these methods to divide them into separate patches, resulting in incomplete semantic information. Thus, we devise an Adaptive patch-word Matching (AdaMatch)

model to adaptively exploit the fine-grained relation between flexible patches and textual words.

2.2 CXR-Report Generation

Medical VLMs are widely used in downstream tasks for chest radiographs, including CXR-to-report generation (Chen et al., 2020b, 2021a; Yang et al., 2021; Wang et al., 2022b; Voutharoja et al., 2023; Yang et al., 2023; Shi et al., 2023; Huang et al., 2023) and report-to-CXR generation (Rombach et al., 2022; Chambon et al., 2022b,a; Lee et al., 2023a,b; Han et al., 2024; Shentu and Al Moubayed, 2024; Hou et al., 2023; Hashmi et al., 2024; Chen et al., 2024) tasks. For CXR-to-report generation, Chen et al., 2021a introduces a cross-modal memory network with shared memory to align images with texts to promote report generation performance. In the report-to-CXR generation task, prior techniques create annotated CXR images from medical reports to augment training data and address privacy concerns, which are categorized into diffusion-based and transformer-based methods. In this paper, we use the cyclic generation (i.e. CXR-to-report and report-to-CXR generation) as use case and evaluation criteria for our fine-grained alignment method (AdaMatch). We also design an AdaMatch-Cyclic to employ AdaMatch to improve the explainability of cyclic generation.

3 Methods

In Fig. 2, we propose an adaptive patch-word matching (AdaMatch) model to associate CXR image regions with words in medical reports, and apply it to CXR-report generation to enhance explainability. Given an input CXR image x^I, the image encoder with several Adaptive Patch extraction modules (AdaPatch) predicts the location and scale of multiple adaptive patches and processes the adaptive patch embeddings z^I. With z^I and text embeddings z^T extracted by text encoder for the medical report x^T, we compute the similarities s^I, s^T between adaptive patches and text tokens to calculate the contrastive loss L_{I2T}, L_{T2I} to optimize AdaMatch. For CXR-to-report generation, we utilize the frozen AdaMatch model to match a predefined textual codebook with the input CXR image x^I to obtain matched keywords k^T and feed a LLM with k^T, the instruction, and image tokens encoded by VQ-GAN from x^I to generate a medical report $x^{T'}$. Then, the AdaMatch model is used to match a predefined visual codebook with the generated report $x^{T'}$ to acquire the matched key-patches k^I. With instruction, $x^{T'}$ and k^I, LLM outputs image tokens of the generated CXR image $x^{I'}$, which is optimized through L_{text} and L_{img}.

3.1 Adaptive Patch-Word Matching (AdaMatch)

Current VLMs (Chen et al., 2020a; Yao et al., 2021) achieve fine-grained alignment between visual objects and textual words, but they may split lung lesions into separate fixed patches due to various sizes of lung lesions. Thus, we propose an Adaptive patch-word Matching (AdaMatch) to locate the important regions, extract adaptive patches for these regions, and align them with corresponding textual words, as shown in Fig. 2 (a).

Image Encoder. To obtain the adaptive patches, the image encoder comprises four stages with feature maps of decreasing scales. Specifically, in the first stage, we first adopt the patch embedding module to split the CXR image $x^I \in \mathbb{R}^{H \times W \times C}$ into N patches with fixed size $s \times s$ and project them to patch embeddings $z^{(i)}(1 \leq i \leq N)$ through a fully connected (FC) layer g, $z^{(i)} = g([a^{(i,1)}; \dots; a^{(i,s \times s)}])$, where $a^{(i,j)}$ indicates the image features for the pixel located at the $q^{(i,j)}$. $q^{(i,j)}$ indicates the coordinates for the image pixel. $[\cdot]$ represents the concatenation operation among the features. Afterward, we pass the patch em-

beddings $z^{(i)}$ with the position embeddings e_{pos} into a transformer encoder R to obtain the outputs, $z^{(i)} = R(z^{(i)} + e_{pos})$. To locate the potential lung lesions, we devise an **Adaptive Patch Extraction module (AdaPatch)** for the rest of stages. In AdaPatch, patch embeddings $z^{(i)}$ are fed into an FC layer, followed by four separate FC layers f_1, f_2, f_3, f_4 to predict offsets $(\delta_x^{(i)}, \delta_y^{(i)})$ and patch sizes $(s_w^{(i)}, s_h^{(i)})$ for adaptive patches, with offsets indicating shifts to the center $(c_x^{(i)}, c_y^{(i)})$ of fixed patches,

$$\delta_x^{(i)} = Tanh(f_1(z^{(i)})), \delta_y^{(i)} = Tanh(f_2(z^{(i)})), \quad (1)$$

$$s_w^{(i)} = ReLU(Tanh(f_3(z^{(i)}))), s_h^{(i)} = ReLU(Tanh(f_4(z^{(i)}))). \quad (2)$$

With the offset and patch size, we compute the position of left-top $(a_x^{(i)}, a_y^{(i)})$ and right-bottom corners $(b_x^{(i)}, b_y^{(i)})$ for each adaptive patch,

$$a_x^{(i)} = c_x^{(i)} + \delta_x^{(i)} - \frac{s_w^{(i)}}{2}, a_y^{(i)} = c_y^{(i)} + \delta_y^{(i)} - \frac{s_h^{(i)}}{2}, \quad (3)$$

$$b_x^{(i)} = c_x^{(i)} + \delta_x^{(i)} + \frac{s_w^{(i)}}{2}, b_y^{(i)} = c_y^{(i)} + \delta_y^{(i)} + \frac{s_h^{(i)}}{2}, \quad (4)$$

and uniformly sample $m \times m$ feature points inside the patches. Since the coordinates may be fractional, bilinear interpolation is used to obtain the sampled feature points $\{\hat{p}^{(j)}\}_{1 \leq j \leq m \times m}$. The embeddings of all sampled points are flattened and fed into an FC layer f_5 to obtain patch embeddings, $z^{(i)} = f_5([\hat{p}^{(i,1)}; \dots; \hat{p}^{(i,m \times m)}])$. Finally, we pass $z^{(i)}$ with position embeddings e_{pos} into a transformer encoder R. The final adaptive patch embeddings $z^I \in \mathbb{R}^{N \times d}$ are the ensemble of $z^{(i)}(1 \leq i \leq N)$, $z^I = E^I(x^I) = \{z_1^I, \dots, z_N^I\}$, where E^I, x^I, and d denote the image encoder, the CXR image, and the dimension of z^I, respectively.

Text Encoder. We adopt a pre-trained text encoder E^T to encode the medical report x^T into text embeddings z^T. To make z^T with the same dimension as z^I, we feed an FC layer with z^T to reduce its dimension, $z^T = E^T(x^T) = \{z_1^T, \dots, z_K^T\}$ ($z^T \in \mathbb{R}^{K \times d}$), where K is the number of text tokens.

Patch-Word Alignment. To exploit the relation between the adaptive patches and textual tokens, we perform fine-grained contrastive representation learning to achieve patch-word alignment. Concretely, for the i-th CXR image x_i^I and j-th medical report x_j^T, we first compute the similarities between all the adaptive patch embeddings $z_n^I(1 \leq n \leq N)$ and all the text embeddings

$z_k^T (1 \leq k \leq K)$, and use the largest similarity $\max_{(1 \leq k \leq K)} (z_n^I)^\top z_k^T$ as the patch-word maximum similarity for n-th adaptive patch embedding. Then, the patch-word maximum similarities for all the adaptive patch embeddings are averaged as the similarity $s_{i,j}^I$ of the i-th CXR image to the j-th medical report, $s_{i,j}^I = \frac{1}{N} \sum_{n=1}^{N} (z_n^I)^\top z_{m_n^I}^T$, where $m_n^I = \arg\max_{(1 \leq k \leq K)} (z_n^I)^\top z_k^T$. Similarly, the similarity of the j-th medical report to the i-th CXR image is defined as, $s_{i,j}^T = \frac{1}{K} \sum_{k=1}^{K} (z_{m_k^T}^I)^\top z_k^T$, where $m_k^T = \arg\max_{(1 \leq n \leq N)} (z_n^I)^\top z_k^T$. We exclude the padded textual tokens when computing the similarity.

With the cross-modal similarities $s_{i,j}^I$ and $s_{i,j}^T$ for the i-th CXR image and j-th medical report, we compute the adaptive patch-word contrastive loss L_i^I for i-th CXR image x_i^I,

$$L_i^{I2T}(x_i^I, \{x_j^T\}_{j=1}^b) = -\frac{1}{b} \log \frac{exp(s_{i,i}^I/\tau)}{\sum_j exp(s_{i,j}^I/\tau)},$$
(5)

where b, τ, and $\{x_i^I, x_i^T\}$ represent the batch size, temperature hyperparameter and the positive CXR-report pair, respectively. Similarly, the adaptive word-patch contrastive loss L_i^T for i-th medical report x_i^T is formulated as,

$$L_i^{T2I}(x_i^T, \{x_j^I\}_{j=1}^b) = -\frac{1}{b} \log \frac{exp(s_{i,i}^T/\tau)}{\sum_j exp(s_{j,i}^T/\tau)}.$$
(6)

The final contrastive loss for a mini-batch is calculated by, $L = \frac{1}{2} \sum_{i=1}^b (L_i^{I2T} + L_i^{T2I})$. With L, AdaMatch learns to locate important patches with varied sizes and exploits the fine-grained relation.

3.2 AdaMatch-based LLM for Cyclic CXR-Report Generation (AdaMatch-Cyclic)

To provide explicit explainability for CXR-report generation task, we propose an AdaMatch-based bidirectional LLM for the cyclic CXR-report generation (AdaMatch-Cyclic) by locating the potential lesions in CXR images and visualizing the appearance of description in reports to guide the generation process, as depicted in Fig. 2 (b).

3.2.1 CXR-to-Report Generation

In CXR-to-report generation, we use AdaMatch to match a predefined textual codebook with the CXR


```
Below is an instruction that describes a task,
paired with an input that provides further
context. Write a response that appropriately
completes the request.
### Instruction: Utilize the entered chest X-ray
images to generate comprehensive free-text
radiology reports.
### Input: 66, 260, 379, 555, 304, ..., 905
### Keywords: Pleural effusions, tube
### Response: A pigtail catheter is placed in the
lower right chest ......
```

Figure 3: The example of instruction data for CXR-to-report generation.

image, providing keywords to guide LLM.

Building textual codebook. Specifically, we first use a pre-trained BioEN (Raza et al., 2022) to extract the related entities from medical reports in the training set, where the related entities are divided into four entity groups, i.e., biological structure, detailed description, disease disorder, and sign symptom. Next, we compute the frequency of each entity and pick top κ_0 entities for each entity group as keywords in the textual codebook.

Keywords Extraction. With the textual codebook, we employ the frozen AdaMatch model to match keywords from the textual codebook with the adaptive patches of CXR images, and obtain a patch-word similarity matrix $s^P \in \mathbb{R}^{M \times N}$ between the keyword tokens and adaptive patch embeddings, where M and N denote the number of keyword tokens and adaptive patches. To extract the most matched keywords for each adaptive patch, we rank the patch-word similarity along the dimension of keyword tokens, obtain the top κ_1 patch-word similarities for each adaptive patch, and extract corresponding keywords k^T. The matched keywords can explain potential lesions in each adaptive patch, to assist LLM in generating medical reports.

Instruction tuning LLM. After keywords extraction, we use a frozen VQ-GAN (Esser et al., 2021) encoder E to encode the input CXR image x^I to the quantized image latent vectors as image tokens $E(x^I)$. For CXR-to-report generation, we adopt the dolly-v2-3b (Conover et al., 2023) model as the pre-trained LLM, and convert CXR-to-report generation dataset in instruction-following format. An example of instruction data for CXR-to-report generation is depicted in Fig. 3, with the instruction, input (the image tokens of the input CXR image), keywords (extracted by AdaMatch), and response parts (the ground-truth medical report). During training, the LLM learns to generate the hidden response part in an autoregressive manner.

By adopting AdaMatch-Cyclic, we can locate potential lesions and interpret them with keywords to enhance explainability.

3.2.2 Report-to-CXR Generation

In report-to-CXR generation, AdaMatch matches a predefined visual codebook with generated reports $x^{T'}$, providing keypatches as guidance for the LLM to synthesize CXR images, where keypatches are important image patches related to the reports.

Building visual codebook. Concretely, we first construct a visual codebook with the most common adaptive patches as keypatches. To collect the most common adaptive patches for CXR images in the training set, AdaMatch matches the adaptive patches of CXR images with textual tokens of medical reports to obtain top κ_2 CXR-report pairs with the highest report-to-CXR similarities s^T. For each CXR-report pair, we compute the word-patch maximum similarity $\max_{(1 \le n \le N)} (z_n^I)^\top z_k^T$ for each textual token, rank the word-patch maximum similarities, and extract the adaptive patches for top κ_3 similarities as keypatches in the visual codebook. Each keypatch includes its adaptive patch and the corresponding features.

Keypatches Extraction. With the visual codebook, the frozen AdaMatch matches the features of keypatches in the visual codebook with textual tokens of the generated report to acquire the word-patch similarity matrix $s^W \in \mathbb{R}^{(\kappa_2 \times \kappa_3) \times K}$, where K is the number of textual tokens. To obtain keypatches related to the generated report, we rank the word-patch similarity along the dimension of keypatches, obtain the top κ_4 word-patch similarity for each textual token, and extract the features of corresponding keypatches k^I.

Instruction tuning LLM. After keypatches extraction, we use a frozen VQ-GAN encoder to convert the matched keypatches k^I into image tokens $E(k^I)$, and feed the LLM with the instruction, generated report, and image tokens of keypatches in the instruction-following format, as shown in Fig. 4. Then, LLM predicts image tokens, decoded by the VQ-GAN into the generated CXR image $x^{I'}$.

AdaMatch-Cyclic allows us to interpret generated reports with matched keypatches, thereby providing explainability to the generation procedure.

3.2.3 Overall Objective

To optimize AdaMatch-Cyclic, we use the standard language modeling objective for both CXR-to-report and report-to-CXR generation. For CXR-

```
Below is an instruction that describes a task,
paired with an input that provides further
context. Write a response that appropriately
completes the request.
### Instruction: Create an AP chest X-
ray image that matches the free-
text radiology reports.
### Input: A pigtail catheter is placed in
the lower right chest ......
### Keypatches: 122, 680, 978
### Response: 719, 421, 551, 421, 742,..., 905
```

Figure 4: The example of instruction data for report-to-CXR generation.

to-report generation, LLM receives the instruction I^t, image tokens of input CXR image $E(x^I)$, and matched keywords k^T, aiming to generate the ground-truth medical report of response part autoregressively. We compute the conditional probability of the k-th token, $P(u_k) = \mathcal{LLM}(I^t, E(x^I), k^T)$, where k is the token index after the response key (### Response:). The report generation loss for the response area is calculated by, $L_{text} = \sum_{i=k}^n -\log P(u_i | u_1, u_2, \cdots, u_{i-1})$, where $[u_1, u_2, \cdots, u_{i-1}]$ denotes the tokenized texts before the response part, and n represents the maximum length of output tokens. Similarly, for report-to-CXR generation, we feed the LLM with the instruction I^v, generated report $x^{T'}$, and image tokens of keypatches $E(k^I)$, and obtain the conditional probability of k-th token w_k, $P(w_k) = \mathcal{LLM}(I^v, x^{T'}, E(k^I))$. The CXR generation loss is defined as: $L_{img} = \sum_{i=k}^n -\log P(w_i | w_1, w_2, \cdots, w_{i-1})$. With L_{text} and L_{img}, LLM can implement bidirectional CXR and report generation according to instructions.

4 Experiments

4.1 Experiment Setting

Datasets. We experiment on two main publicly available chest X-ray datasets, i.e. MIMIC-CXR (Johnson et al., 2019) and OpenI (Demner-Fushman et al., 2016) datasets. **MIMIC-CXR** dataset comprises 473,057 images and 206,563 reports from 63,478 patients. We use the official splits, i.e. 368,960 for training, 2,991 for validation, and 5,159 for testing. **OpenI** contains 3,684 report-image pairs with 2,912 for training and 772 for testing. Unlike prior methods, we utilize both finding and impression sections as medical reports. **Implementation Details.** In the AdaMatch-Cyclic model, we first train the AdaMatch model and then use the frozen AdaMatch to train LLM. We adopt VQ-GAN (Esser et al., 2021) models pre-trained on

Table 1: Comparison of CXR-to-report generation performance on the MIMIC-CXR and the OpenI datasets.

Methods	MIMIC-CXR						OpenI					
	B-1	B-2	B-3	B-4	M	R-L	B-1	B-2	B-3	B-4	M	R-L
R2Gen	0.3553	0.2232	0.1523	0.1038	0.1412	0.2784	0.3992	0.2407	0.1518	0.0973	0.1390	0.3052
R2GenCMN	0.3719	0.2332	0.1538	0.1053	0.1501	0.2827	0.4091	0.2493	0.1594	0.1045	0.1509	0.3181
Joint-TriNet	0.3585	0.2266	**0.1550**	0.1021	0.1425	0.2788	0.3833	0.2409	0.1598	0.1078	0.1457	0.3293
XProNet	0.3532	0.2212	0.1498	0.1052	0.1415	0.2811	0.4114	0.2502	0.1598	0.1045	0.1457	0.3240
ITHN	0.3623	0.2128	0.1402	0.0992	0.1488	0.2622	0.2661	0.1516	0.0976	0.0663	0.1561	0.2617
M2KT	0.3661	0.2192	0.1465	0.1044	0.1528	0.2673	0.2559	0.1381	0.0819	0.0523	0.1468	0.2439
AdaMatch-Cyclic	**0.3793**	**0.2346**	0.1540	**0.1060**	**0.1625**	**0.2859**	**0.4161**	**0.3002**	**0.2073**	**0.1446**	**0.1621**	**0.3656**

CXR Image	Ground-truth	M2KT	AdaMatch-Cyclic	Keywords
	PA and lateral views of the chest. Bilateral upper lobe scarring is seen with superior retraction of the hila. The lung volumes are relatively low. There is no evidence of superimposed acute process. Cardiomediastinal silhouette is stable. Surgical clips in the upper abdomen again noted. Osseous structures are essentially unremarkable noting probable right glenoid orthopedic hardware. Bilateral upper lobe scarring unchanged without evidence of superimposed acute process.	PA and lateral views of the chest were obtained. the patient is status post median sternotomy and cabg. The heart is normal in size. the mediastinal and hilar contours appear unchanged. There is no pleural effusion or pneumothorax. The pulmonary vasculature is not engorged. Patchy opacity in the right lower lung is concerning for pneumonia in the appropriate clinical setting. Patchy opacities in the right lung base may reflect atelectasis or pneumonia.	Heart size is mildly enlarged. Mediastinal and hilar contours are unremarkable. Pulmonary vasculature is normal. Mild elevation of the right hemidiaphragm is unchanged. Scarring within the left lung base is re-demonstrated with a left basilar calcified granuloma re-demonstrated. There is no focal consolidation pleural effusion or pneumothorax. No acute osseous abnormality is visualized. Clips from prior cholecystectomy are noted in the right upper quadrant of the abdomen. A percutaneous catheter is noted within the right upper quadrant of the abdomen. No evidence for pneumonia or congestive heart failure.	**upper lobe**, **bilateral**, **scarring**, thoracic vertebral body, edema, reticular opacities, semi-upright, **tube** side port, right pneumotho rax
CXR Image	Ground-truth	R2GenCMN	AdaMatch-Cyclic	Keywords
	Small calcification right lung base with appearance of old granulomatous disease. Also small perihilar calcified lymph XXXX. Lungs are clear. No active parenchymal disease. No XXXX of pleural effusions. No pulmonary edema. Normal heart size. No XXXX of active cardiopulmonary disease. Unchanged.	Heart size is normal. The mediastinal and hilar contours are normal. The pulmonary vasculature is clear. Lungs are clear. No pleural effusion or pneumothorax is seen. There are no acute osseous abnormalities. No acute cardiopulmonary abnormality.	The trachea is midline. The heart is normal in size. The mediastinum is unremarkable. Mild granulomatous sequela are noted. The lungs are grossly clear. There is no pneumothorax. No acute disease.	lung, old, **calcification**, lymph, **perihilar**, airway, indwelling, **calcified granuloma**, hypoinflation, hiatal

Figure 5: Qualitative comparison of CXR-to-report generation on the MIMIC-CXR (1st row) and the OpenI (2nd row) datasets, highlighting similar meanings in colored text. The keywords are obtained from AdaMatch.

Table 2: Performance of CXR-to-report generation compared to GPT-4V on 50 selected MIMIC-CXR cases.

Methods	B-1	B-2	B-3	B-4
R2Gen	0.3513	0.2174	0.1447	0.1025
R2GenCMN	0.3587	0.2200	0.1436	0.0969
Joint-TriNet	0.3596	0.2218	0.1481	0.1026
XProNet	0.3356	0.2071	0.1374	0.0941
ITHN	0.3301	0.1839	0.1121	0.0723
M2KT	0.3626	0.2123	0.1391	0.0957
GPT-4V	0.2275	0.0878	0.0378	0.0166
AdaMatch-Cyclic	**0.3754**	**0.2303**	**0.1520**	**0.1058**

Table 3: Comparison of report-to-CXR generation performance on the MIMIC-CXR and the OpenI datasets.

Methods	MIMIC-CXR		OpenI	
	FID↓	NIQE↓	FID↓	NIQE↓
Stable diffusion	9.2334	3.7894	8.2946	6.3496
Adapting-Med	8.2758	3.8871	5.8557	4.6534
RoentGen	9.5411	3.8834	6.5675	4.9085
UniXGen	6.7212	3.7125	11.9890	4.6610
LLM-CXR	2.1788	3.5969	1.6597	3.8206
AdaMatch-Cyclic	**1.0916**	**3.3931**	**1.5938**	**3.3096**

the MIMIC-CXR and OpenI datasets, respectively, and the dolly-v2-3b (Conover et al., 2023) model as pre-trained LLM. The source code will be released. Please see appendix C for more details.

Evaluation Metrics. We assess CXR-to-report generation using BLEU (B), METEOR (M), and ROUGE-L (R-L) (Chen et al., 2020b), and report-to-CXR generation using FID (Heusel et al., 2017) and NIQE (Mittal et al., 2012). The retrieval performance between CXR and report is assessed through the exact report in the top K retrieved reports for a given CXR image (R@K, K={1, 5, 10}).

4.2 Comparison with State-of-the-Arts.

4.2.1 CXR-to-Report Generation

We compare AdaMatch-Cyclic with current CXR-to-report generation methods on the MIMIC-CXR and the OpenI datasets, including R2Gen (Chen et al., 2020b), R2GenCMN (Chen et al., 2021a), Joint-TriNet (Yang et al., 2021), XProNet (Wang et al., 2022b), ITHN (Voutharoja et al., 2023), and M2KT (Yang et al., 2023). Since these methods are mainly trained on the finding section, we reimplement them on both the finding and impression sections. In Table 1, AdaMatch-Cyclic achieves the best performance of 0.3793 in BLEU-1 on the MIMIC-CXR dataset, with superior generalization
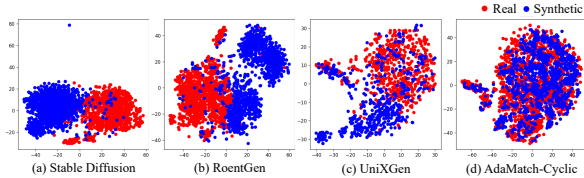
Figure 6: The t-SNE visualization of the real and synthetic CXR images on the MIMIC-CXR dataset.
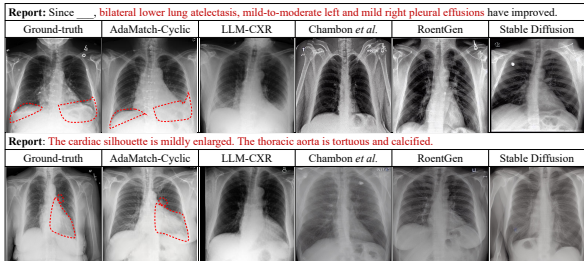

Figure 7: Generated CXR images of the MIMIC-CXR (1st row) and OpenI (2nd row) datasets with highlighted regions.

Table 4: Comparison of CXR-report retrieval performance (%) on MIMIC-CXR dataset.

Methods	CXR-to-Report			Report-to-CXR		
	R@1	R@5	R@10	R@1	R@5	R@10
Fang et al. (2015)	18.60	43.10	56.10	18.13	43.20	55.97
Chauhan et al. (2020)	5.37	19.43	30.73	5.40	20.23	30.23
ConVIRT (Zhang et al., 2022)	30.10	53.90	63.80	29.20	54.70	64.40
GLoRIA (Huang et al., 2021)	30.30	57.50	66.50	24.00	51.80	62.80
JoImTeR-Net (Ji et al., 2021)	18.93	46.20	58.67	19.07	45.27	58.50
MGCA (Wang et al., 2022a)	25.80	51.90	62.10	27.90	51.20	61.60
LIMITR (Dawidowicz et al., 2023)	39.70	63.20	71.70	37.70	62.10	71.30
Motor (Lin et al., 2023)	10.96	31.93	42.90	12.00	33.10	44.32
AdaMatch	**51.47**	**86.19**	**94.77**	**51.18**	**86.46**	**94.60**

Table 5: Ablation study on AdaMatch w/o AdaPatch.

AdaPatch	CXR-to-Report			Report-to-CXR		
	R@1	R@5	R@10	R@1	R@5	R@10
✗	48.77	83.89	92.94	48.72	83.95	92.90
✓	**51.47**	**86.19**	**94.77**	**51.18**	**86.46**	**94.60**

to OpenI dataset using the same model. Compared to GPT-4V, AdaMatch-Cyclic significantly outperforms in BLEU-1 by 0.1479, as listed in Table 2.

In Fig. 5, we compare AdaMatch-Cyclic's performance with M2KT and R2GenCMN. AdaMatch-Cyclic accurately captures relevant keywords like 'scarring', and describes 'osseous abnormality' and 'clips in the right upper quadrant of the abdomen' in the first case, showcasing superior report quality over previous methods.

4.2.2 Report-to-CXR Generation

We quantitatively compare our method with text-to-image generation method like Stable Diffusion (Rombach et al., 2022) and report-to-CXR generation methods, such as Adapting-Med (Chambon et al., 2022b), RoentGen (Chambon et al., 2022a), UniXGen (Lee et al., 2023a), and LLM-CXR (Lee et al., 2023b). In Table 3, AdaMatch-Cyclic achieves the highest FID scores on both datasets, showing its superior effectiveness in generating CXR. To compare the high-level feature distribution of CXR images generated by different methods, we randomly select 1,000 cases from test set, and apply t-SNE visualization to the real and synthetic CXR images on the MIMIC-CXR dataset. In Fig. 6, while existing methods' synthetic CXR images differ from real ones, AdaMatch-Cyclic's almost overlap with real ones, indicating its superiority in report-to-CXR generation.

In Fig. 7, we display generated CXR images from the MIMIC-CXR and OpenI datasets. In the

first example, AdaMatch-Cyclic accurately synthesizes 'left and right pleural effusions', while other methods fail to generate such features. This suggests the superior ability of our method to produce realistic CXR images based on input reports.

4.2.3 CXR-Report Retrieval

To demonstrate the superiority of AdaMatch, we evaluate the CXR-to-report and report-to-CXR retrieval performance on the MIMIC-CXR dataset compared to existing methods. We utilize the AdaMatch to compute the similarities between CXR images and reports, rank the similarities, and obtain the retrieval results. As shown in Table 4, AdaMatch significantly outperforms LIMITR with R@1 of 11.77% and 13.48% for CXR-to-report and report-to-CXR retrieval. This indicates the superior ability of our method to extract distinct semantic features and align them accurately between CXR images and medical reports.

In Table 5, to ablate AdaPatch, we remove the layers to predict offset and scale. We then aligned grid features from the last stage of the pyramid vision Transformer with textual tokens to compute the image-text similarity for retrieval tasks.

4.3 Ablation Study

Effectiveness of AdaPatch. We evaluate the effectiveness of AdaPatch by comparing the CXR-report retrieval performance of AdaMatch with and without AdaPatch. To ablate AdaPatch, we remove the layers to predict offset and scale, and compute the similarity between grid features of the last stage and textual tokens for retrieval tasks.

Table 6: Effectiveness of report-to-CXR and CXR-to-report generation tasks.

Report-to-CXR	CXR-to-Report	B-1↑	B-2↑	B-3↑	B-4↑	M↑	R-L↑	FID↓	NIQE↓
	✓	0.3542	0.1973	0.1190	0.0758	0.1329	0.2392	-	-
✓		-	-	-	-	-	-	1.7128	4.0391
✓	✓	0.3793	0.2346	0.1540	0.1060	0.1625	0.2813	1.0916	3.3931

Table 7: Comparison of AdaPatch and other adaptive vision models.

Models	CXR-to-Report			Report-to-CXR		
	R@1	R@5	R@10	R@1	R@5	R@10
A-ViT	49.78	83.14	92.77	49.49	83.19	92.94
AdaViT	50.14	84.52	93.20	50.54	84.10	93.83
AdaPatch	**51.47**	**86.19**	**94.77**	**51.18**	**86.46**	**94.60**

Table 8: CXR-report retrieval performance of AdaMatch with different stages.

Stage	CXR-to-Report			Report-to-CXR		
	R@1	R@5	R@10	R@1	R@5	R@10
2	42.28	78.75	89.85	42.81	78.96	89.67
3	**51.47**	**86.19**	**94.77**	**51.18**	**86.46**	**94.60**
4	49.68	83.04	92.67	49.39	83.09	92.84

In Table 5, AdaPatch significantly improves R@1 by approximately 3% for both retrieval directions, highlighting its effectiveness. To demonstrate the optimality of the AdaPatch design, we compare AdaPatch with other adaptive vision models, such as A-ViT (Yin et al., 2022) and AdaViT (Meng et al., 2022), in the CXR-report retrieval task. As shown in Table 7, AdaPatch exhibits superior retrieval performance compared with other methods.

Effectiveness of Cyclic Generation. To validate the necessity of report-to-CXR and CXR-to-report generation tasks, we remove each task and evaluate their performance, as listed in Table 6. When ablating report-to-CXR generation task, the performance of CXR-to-report generation decreases substantially compared to our method, and vice versa. These suggest that the CXR-to-report and report-to-CXR generation can indeed benefit each other, and their interaction is crucial for the overall performance of the model.

Image Encoder with Different Stages. In AdaMatch, the image encoder comprises several stages, with AdaPatch modules in the stages. To assess the effectiveness of the stage number, we compare the retrieval performance of AdaMatch with two, three, or four stages. Table 8 shows that the three-stage image encoder achieves the highest R@1 of 51.47%, indicating that the features from
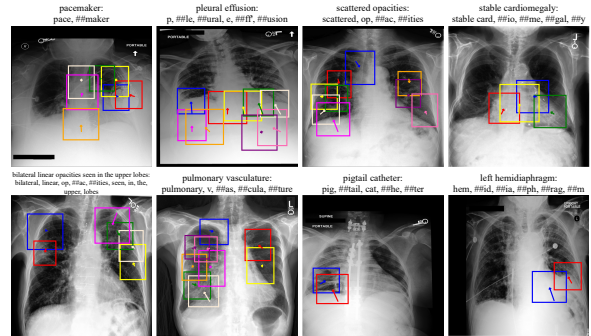

Figure 8: Visualization of texts and its adaptive patches. Boxes and arrows in different colors show adaptive patches and their center shifts from fixed patches.

the 3rd stage are most beneficial for CXR-report retrieval.

Visualization of Adaptive Patches. We visualize adaptive patches in AdaPatch of some examples from the MIMIC-CXR dataset in Fig. 8. Each example includes text, its tokens, and the corresponding CXR image with marked adaptive patches. The patches are represented by colored bounding boxes with arrows indicating the center shift from the fixed patch. In the first example, adaptive patches cover the pacemaker and wire, and the second one can find pleural effusions, implying the capacity of AdaPatch to localize regions relevant to input text tokens accurately.

5 Conclusion

We propose AdaMatch for fine-grained image-text alignment, which presents the first work to adaptively associate image patches with words and improve the explainability of cyclic CXR-report generation. It includes AdaPatch to acquire adaptive patches for abnormal regions and performs patch-word alignment between adaptive patches with textual tokens. We implement cyclic CXR-report generation by using AdaMatch to provide explanations for the generation process. In addition, the fine-grained cyclic generation process provides a natural explainability for the alignment between CXR images and reports. Extensive experiments on two CXR datasets show the effectiveness of our method and its superiority over previous methods.

Limitation

While our research has made significant strides in utilizing chest X-ray datasets, it is important to acknowledge certain limitations. Our experiments predominantly focus on chest X-ray datasets due to their availability of large-scale images paired with high-quality medical reports. However, these datasets primarily consist of patients from the ICU, potentially skewing our model towards severe disease domains. In our future endeavors, we intend to expand the scope of our methodology by applying the proposed AdaMatch-Cyclic to multi-domain scenarios, thereby mitigating this limitation and enhancing the versatility of our approach.

Acknowledge

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. pages 72–78.

Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. 2022a. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*.

Pierre Chambon, Christian Bluethgen, Curtis P Langlotz, and Akshay Chaudhari. 2022b. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*.

Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. 2020. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *MICCAI*, pages 529–539. Springer.

Wenting Chen, Pengyu Wang, Hui Ren, Lichao Sun, Quanzheng Li, Yixuan Yuan, and Xiang Li. 2024. Medical image synthesis via fine-grained image-text alignment and anatomy-pathology prompting. *arXiv preprint arXiv:2403.06835*.

Xiaofei Chen, Yuting He, Cheng Xue, Rongjun Ge, Shuo Li, and Guanyu Yang. 2023. Knowledge boosting: Rethinking medical contrastive vision-language pre-training. In *MICCAI*, pages 405–415. Springer.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020a. Uniter: Universal image-text representation learning. In *ECCV*, pages 104–120. Springer.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021a. Cross-modal memory networks for radiology report generation. In *ACL*, pages 5904–5914.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020b. Generating radiology reports via memory-driven transformer. In *EMNLP*, pages 1439–1449.

Zhiyang Chen, Yousong Zhu, Chaoyang Zhao, Guosheng Hu, Wei Zeng, Jinqiao Wang, and Ming Tang. 2021b. Dpt: Deformable patch-based transformer for visual recognition. In *ACM MM*, pages 2899–2907.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Gefen Dawidowicz, Elad Hirsch, and Ayellet Tal. 2023. Limitr: Leveraging local information for medical image-text representation. *arXiv preprint arXiv:2303.11755*.

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. *JAMIA*, 23(2):304–310.

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming transformers for high-resolution image synthesis. In *CVPR*, pages 12873–12883.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *CVPR*, pages 1473–1482.

Woojung Han, Chanyoung Kim, Dayun Ju, Yumin Shim, and Seong Jae Hwang. 2024. Advancing text-driven chest x-ray generation with policy-based reinforcement learning. *arXiv preprint arXiv:2403.06516*.

Anees Ur Rehman Hashmi, Ibrahim Almakky, Mohammad Areeb Qazi, Santosh Sanjeev, Vijay Ram Papineni, Dwarikanath Mahapatra, and Mohammad Yaqub. 2024. Xreal: Realistic anatomy and pathology-aware x-ray generation via controllable diffusion model. *arXiv preprint arXiv:2403.09240*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30:6629–6640.

Zeyi Hou, Ruixin Yan, Qizheng Wang, Ning Lang, and Xiuzhuang Zhou. 2023. Diversity-preserving chest radiographs generation from reports in one stage. In *MICCAI*, pages 482–492. Springer.

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 *ICCV*, pages 3942–3951.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In *CVPR*, pages 19809–19818.

Jiho Jang, Chaerin Kong, Donghyeon Jeon, Seonhoon Kim, and Nojun Kwak. 2023. Unifying vision-language representation space with single-tower transformer. In *AAAI*, volume 37, pages 980–988.

Zhanghexuan Ji, Mohammad Abuzar Shaikh, Dana Moukheiber, Sargur N Srihari, Yifan Peng, and Mingchen Gao. 2021. Improving joint learning of chest x-ray and radiology report by word region alignment. In *MLMI*, page 110–119. Springer.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR.

Chaoya Jiang, Haiyang Xu, Wei Ye, Qinghao Ye, Chenliang Li, Ming Yan, Bin Bi, Shikun Zhang, Fei Huang, and Ji Zhang. 2023. Copa: Efficient vision-language pre-training through collaborative object- and patch-text alignment. In *ACM MM*, pages 4480–4491.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594. PMLR.

Hyungyung Lee, Wonjae Kim, Jin-Hwa Kim, Tackeun Kim, Jihang Kim, Leonard Sunwoo, and Edward Choi. 2023a. Unified chest x-ray and radiology report generation model with multi-view chest x-rays. *arXiv preprint arXiv:2302.12172*.

Suhyeon Lee, Won Jun Kim, and Jong Chul Ye. 2023b. Llm itself can read and generate cxr images. *arXiv preprint arXiv:2305.11490*.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. 2020a. Unimo: Towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, pages 121–137. Springer.

Bingqian Lin, Zicong Chen, Mingjie Li, Haokun Lin, Hang Xu, Yi Zhu, Jianzhuang Liu, Wenjia Cai, Lei Yang, Shen Zhao, et al. 2023. Towards medical artificial general intelligence via knowledge-enhanced multimodal pretraining. *arXiv preprint arXiv:2304.14204*.

Che Liu, Sibo Cheng, Miaojing Shi, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2023. Imitate: Clinical prior guided hierarchical vision-language pre-training. *arXiv preprint arXiv:2310.07355*.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *ICLR*.

Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser-Nam Lim. 2022. Adavit: Adaptive vision transformers for efficient image recognition. In *CVPR*, pages 12309–12318.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.*, 20(3):209–212.

Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. 2022. Multimodal understanding and generation for medical images and text via vision-language pre-training. *IEEE J. Biomed. Health Inform.*, 26(12):6070–6080.

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 *ICML*, pages 8748–8763. PMLR.

Shaina Raza, Deepak John Reji, Femi Shajan, and Syed Raza Bashir. 2022. Large-scale application of named entity recognition to biomedicine and epidemiology. *PLOS Digit. Health*, 1(12):1–18.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626.

Junjie Shentu and Noura Al Moubayed. 2024. Cxr-irgen: An integrated vision and language model for the generation of clinically accurate chest x-ray image-report pairs. In *WACV*, pages 5212–5221.

Yanzhao Shi, Junzhong Ji, Xiaodan Zhang, Liangqiong Qu, and Ying Liu. 2023. Granularity matters: Pathological graph-driven cross-modal alignment for brain ct report generation. In *EMNLP*.

Bhanu Voutharoja, Lei Wang, and Luping Zhou. 2023. Automatic radiology report generation by learning with increasingly hard negatives. pages 2427–2434.

Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. 2023. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *arXiv preprint arXiv:2305.19894*.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2022a. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *NeurIPS*, 35:33536–33549.

Jun Wang, Abhir Bhalerao, and Yulan He. 2022b. Cross-modal prototype driven network for radiology report generation. In *ECCV*, pages 563–579. Springer.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. 2023. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *CVPR*, pages 19175–19186.

Yahui Xu, Yi Bin, Jiwei Wei, Yang Yang, Guoqing Wang, and Heng Tao Shen. 2023. Multi-modal transformer with global-local alignment for composed query image retrieval. *IEEE Trans. on Multimed.*

Bin Yan and Mingtao Pei. 2022. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *AAAI*, volume 36, pages 2982–2990.

Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *CVPR*, pages 15671–15680.

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. *Med. Image Anal.*, 86:102798.

Yan Yang, Jun Yu, Jian Zhang, Weidong Han, Hanliang Jiang, and Qingming Huang. 2021. Joint embedding of deep visual and semantic features for medical image report generation. *IEEE Trans. on Multimed.*

Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. 2022. A-vit: Adaptive tokens for efficient vision transformer. In *CVPR*, pages 10809–10818.

Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. 2021. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *CVPR*, pages 11782–11791.

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 *MLHC*, pages 2–25. PMLR.

Appendix

Abstract. In this supplementary material, we provide additional information about the proposed method. Appendix A illustrates ablation studies on AdaMatch-Cyclic. Appendix B demonstrates visual results of CXR-to-report generation, report-to-CXR generation, and adaptive patches. Appendix C provides the implementation details of the proposed method.

A   Ablation Studies on AdaMatch-Cyclic

In the proposed AdaMatch-Cyclic, we leverage the AdaMatch to obtain keywords for CXR-to-report generation and keypatches for the report-to-CXR generation. To analyze the effectiveness of keywords and keypatches for the generation process, we ablate them in the AdaMatch-Cyclic on the MIMIC-CXR dataset to evaluate the CXR-report generation performance.

Effectiveness of keywords. As listed in Table 9, we ablate both keywords and keypatches of AdaMatch-Cyclic to create a baseline model. When we employ the keywords in the baseline model, the CXR-to-report generation performance improves significantly by about 0.03 in BLEU-4, indicating the effectiveness of the keywords obtained from AdaMatch. To further analyze the influence of the number of keywords N_w, we train the proposed AdaMatch-Cyclic with different $N_w = \{10, 15, 20\}$. In Table 10, AdaMatch-Cyclic achieves the best report generation performance with the ROUGE-L of 0.2859, when the N_w is set to 10.

Effectiveness of keypatches. In Table 9, when we further apply keypatches to the baseline model with keywords, the CXR generation performance boosts remarkably by about 0.5 in FID score, implying the effectiveness of keypatches provided by AdaMatch. To investigate the influence of the number of keypatches N_p, we analyze the CXR

Table 9: The effectiveness of keywords and keypatches. B-4, M, and R-L represent BLEU-4, METEOR, and ROUGE-L, respectively.

Keywords	Keypatches	B-4	M	R-L	FID
		0.0778	0.1295	0.2402	1.5378
✓		0.1059	0.1649	0.2813	1.5132
✓	✓	0.1060	0.1625	0.2859	1.0916

Table 10: The analysis on the different number of keywords (N_w). B, M, and R-L represent BLEU, METEOR, and ROUGE-L, respectively.

N_w	B-1	B-2	B-3	B-4	M	R-L
10	0.3793	0.2346	0.1540	0.1060	0.1625	0.2859
15	0.3849	0.2265	0.1424	0.0937	0.1502	0.2668
20	0.3769	0.2281	0.1478	0.1001	0.1519	0.2762

Table 11: The analysis on the different number of keypatches(N_p).

N_p	5	10	15
FID	1.0916	1.6544	1.6720

generation performance when the number of keypatches ranges from 5 to 15. As shown in Table 11, AdaMatch-Cyclic achieves the best CXR generation performance with the FID score of 1.0916, when the number of keypatches is 5.

B   Visual Results

To prove the effectiveness of AdaMatch-Cyclic for CXR-to-report and report-to-CXR generation tasks, we demonstrate some visual results for both tasks. Moreover, to further visualize the explanation provided by AdaMatch, i.e. the adaptive patches and the corresponding texts.

CXR-to-Report Generation. As shown in Fig. 9, we compare CXR-to-report generation performance with existing methods on the MIMIC-CXR and OpenI datasets. In the first row, the proposed AdaMatch-Cyclic can capture the 'Feeding tube tip' and 'mild bibasilar atelectasis', while M2KT (Yang et al., 2023) cannot observe such device and lung abnormality. The second case of the OpenI dataset shows that 'cardiac silhouette is mildly enlarged' and 'degenerative changes of the thoracic spine' can be discovered by our method. These imply that our AdaMatch-Cyclic can generate a more comprehensive and complete medical report with the guidance of generated keywords in comparison to current methods.

Report-to-CXR Generation. Fig. 10 visualizes the real and synthetic CXR images of the MIMIC-CXR and OpenI datasets in comparison with existing methods. As depicted in the first example, the CXR image generated by AdaMatch-Cyclic shows 'patchy opacity in the right lung', while the CXR images generated by other methods do not include this pattern. In the second example, our AdaMatch-Cyclic can generate the CXR image

CXR Image	Ground-truth	M2KT	AdaMatch-Cyclic	Keywords
	The nasogastric tube is in adequate position and there is a resolution of the gastric distention. There is still mild bibasilar atelectasis. There are no pneumothorax no pleural effusion. The cardiac and mediastinal contour are unchanged.	Frontal and lateral views of the chest were obtained. the patient is status post median sternotomy and cabg. The heart is mildly enlarged. the mediastinal and hilar contours are unchanged. There is a small right pleural effusion with adjacent atelectasis. There is no pneumothorax. The pulmonary vasculature is not engorged. Patchy opacities in the lung bases likely reflect areas of atelectasis. No focal consolidation pleural effusion or pneumothorax is present.	Feeding tube tip is in the mid stomach. Right hilum is enlarged partially obscured by catheter tubing unchanged. There is more prominent mild bibasilar atelectasis and mild diffuse mild vascular prominence but no alveolar edema. Heart size is normal. Enlarged hila with central venous catheter obscuring hilar structures. Suggest repeat with removal of line or scope.	**nasogastric**, **atelectasis**, hilum, diffuse, nodular opacities, spondylotic, **tube** side port, curvature, emphysematous changes, gaseous, hyperlucency, volume loss, subdiaphragmatic free air, lateral

CXR Image	Ground-truth	R2GenCMN	AdaMatch-Cyclic	Keywords
	There is stable mild enlargement of the cardiac silhouette. Stable mediastinal silhouette. There are low lung volumes with bronchovascular crowding. Scattered xxxx opacities in the right lung base xxxx representing foci of subsegmental atelectasis with scattered airspace opacities in the medial left lower lobe. No pneumothorax. No pleural effusion. Degenerative changes of the thoracic spine possibly consistent with dish. Low lung volumes with mild cardiomegaly and scattered right basilar subsegmental atelectasis and scattered retrocardiac airspace opacities.	PA and lateral views of the chest provided. There is no focal consolidation effusion or pneumothorax. The cardiomediastinal silhouette is normal. Imaged osseous structures are intact. No free air below the right hemidiaphragm is seen. No acute intrathoracic process.	The cardiac silhouette is mildly enlarged. There is central pulmonary vascular congestion with diffusely increased interstitial and mild patchy airspace opacities. There is no pleural effusion or pneumothorax. There is mild degenerative changes of the thoracic spine. Mild cardiomegaly and central pulmonary vascular congestion with mild pulmonary edema.	**cardiac**, scattered, **enlargement**, mediastinal, subsegmental, retrosternal space, right base, diaphragm, unremarkable, thin

Figure 9: Qualitative comparison with existing methods in CXR-to-report generation on the MIMIC-CXR (1st row) and OpenI (2nd row) datasets. The texts in different colors show similar meanings. The keywords on the right are obtained from AdaMatch model.
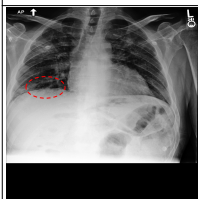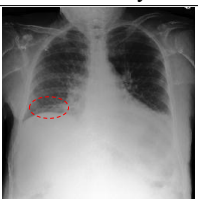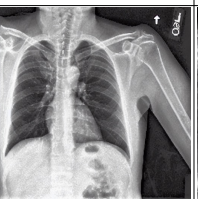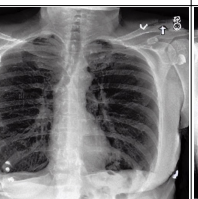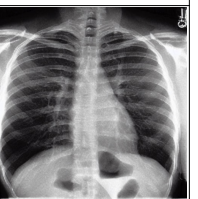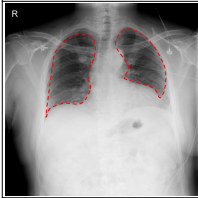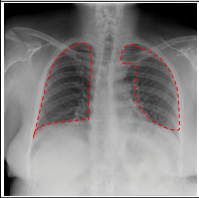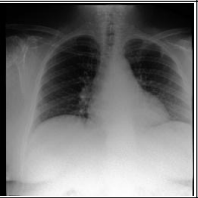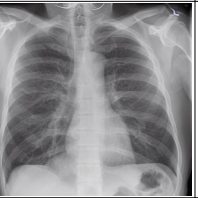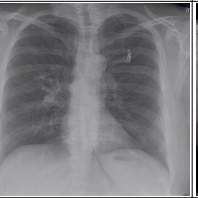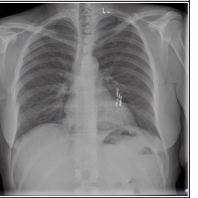
Report: Patchy opacity in the right lung base may reflect atelectasis.					
Ground-truth	AdaMatch-Cyclic	LLM-CXR	Chambon *et al.*	RoentGen	Stable Diffusion

Report: Low lung volumes bilaterally with central bronchovascular crowding.					
Ground-truth	AdaMatch-Cyclic	LLM-CXR	Chambon *et al.*	RoentGen	Stable Diffusion

Figure 10: Visualization of the real and synthetic CXR images of the MIMIC-CXR (1st row) and the OpenI (2nd row) datasets.
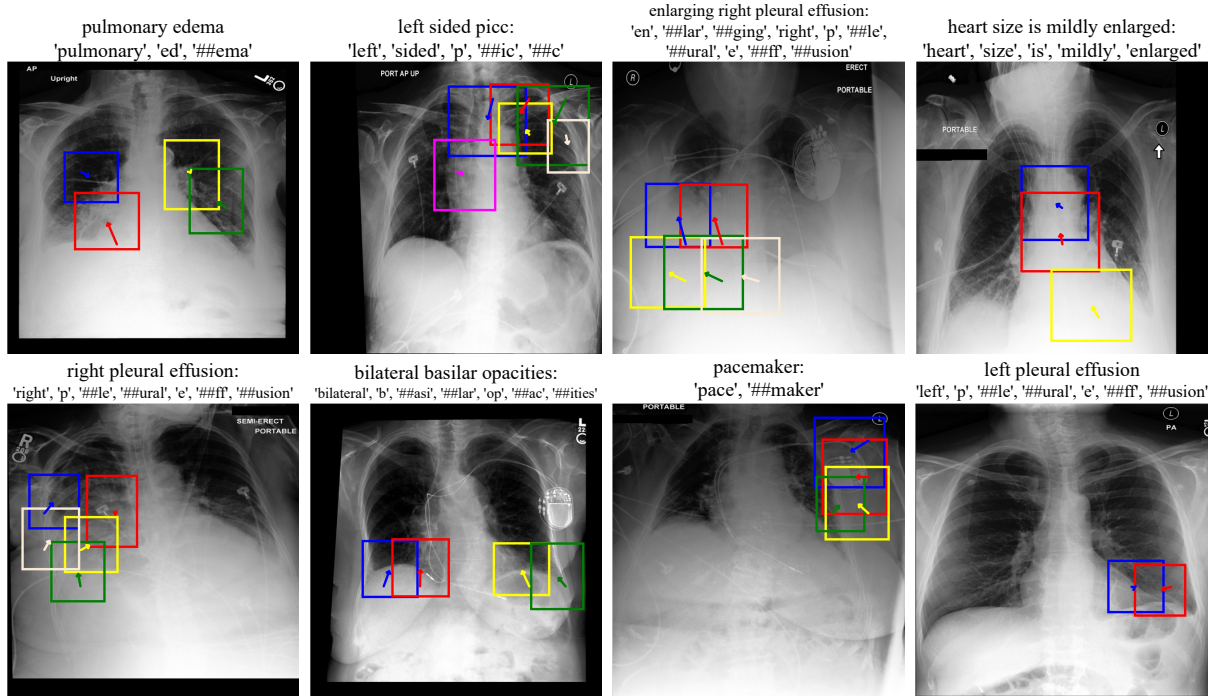
Figure 11: Visualization of texts and the corresponding adaptive patches. The boxes and arrows in different colors show adaptive patches and their center shifts from fixed patches.

with 'low lung volumes'. These indicate the superiority of our AdaMatch-Cyclic over existing methods in report-to-CXR generation.

Adaptive Patches and Texts. In Fig. 11, we visualize adaptive patches in AdaPatch, the textual words, and the textual tokens for different cases from the MIMIC-CXR datasets. We highlight the adaptive patches with bounding boxes in different colors. Each bounding box has an arrow inside to show the shift of center from the fixed patch to the adaptive patch. In the first example, adaptive patches cover the pulmonary edema. Meanwhile, adaptive patches of the second example show the correct position of PICC device in the left lung of the CXR image. These suggest that AdaMatch-Cyclic can show the correspondence between the adaptive patches and textual words to provide the correct explanation for the CXR-report generation.

C   Implementation Details

In the AdaMatch-Cyclic model, we first train the AdaMatch model and then use the frozen AdaMatch to train LLM. The AdaMatch model consists of an image encoder and a text encoder. We utilize the DPT-medium (Chen et al., 2021b) as the image encoder that includes AdaPatch module

in stage 2 and 3. The image encoder is pre-trained on the MIMIC-CXR dataset with the disease classification task. We adopt the pre-trained BioClinicalBERT (Alsentzer et al., 2019) as text encoder. The image and text encoders are followed by two convolutional layers with batch normalization and the ReLU activation function, respectively, to reduce the feature dimension to 256. The patch size s is set to 4 for stage 1, and 2 for stage 2, 3, and 4. The number of sampled feature points m is 3. We optimize the AdaMatch using LAMB optimizer ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-4}$). The cosine learning rate scheduler with the base learning rate (lr_b) of 6×10^{-3} is adopted to linearly warm up to the peak learning rate (i.e. $lr\_p = lr\_b \times \sqrt{\frac{bs\_all}{512}}$) during the first quarter of the total training epochs, where bs_all denotes the effective total batch size. The total training epoch is 15 and the per GPU batch size is 112. In AdaMatch-Cyclic, we use the pre-trained VQ-GAN (Esser et al., 2021) model to encode CXR images into image tokens and decode image tokens into CXR images. We adopt the dolly-v2-3b (Conover et al., 2023) model as the pre-trained LLM. The LLM has 5,1845 token types with the first 5,0821 token types for text tokens and the rest 1,024 token types for image tokens. We add 5,0821 to each image token value encoded by VQ-GAN. To decode the image tokens into images,

we subtract 5,0821 from the image tokens generated by LLM and feed image tokens to VQ-GAN decoder to obtain the generated CXR images. We train the LLM with the AdamW (Loshchilov and Hutter, 2018) optimizer. The learning rate is initialized as 5×10^{-6} and the total training epoch is 5. The per GPU batch size is set to 24. The hyper-parameters $\kappa_0, \kappa_1, \kappa_2, \kappa_3, \kappa_4$ are set as 200, 10, 1000, 20, and 5, respectively. In CXR-to-report generation, we extract keywords for each adaptive patch and use the keywords with the top 10 patch-word similarities as hints of LLM. In addition, we use 5 keypatches to guide the LLM in report-to-CXR generation. All the experiments are conducted on 8 Nvidia A100 40GB GPUs. There is no overlap of patients among different subsets. We discard the medical reports with fewer than 3 tokens from both datasets. All the CXR images with different sizes are resized to 256×256 pixels.