# AHIVE: Anatomy-aware Hierarchical Vision Encoding for Interactive Radiology Report Retrieval

Sixing Yan[1,2,5]*, William K. Cheung[1], Ivor W. Tsang[2,3,4], Keith Chiu[5], Terence M. Tong[6],
Ka Chun Cheung[1,7], and Simon See[7]
[1]Hong Kong Baptist University, [2]CFAR and IHPC, Agency for Science, Technology
and Research, Singapore, [3]SCSE, Nanyang Technological University,
[4]AAII, University of Technology Sydney, [5]Queen Elizabeth and Kwong Wah Hospitals, Hong Kong,
[6]Tuen Mun Hospital, Hong Kong, [7]NVIDIA AI Technology Center, NVIDIA Corporation
{cssxyan, william}@comp.hkbu.edu.hk, ivor_tsang@cfar.a-star.edu.sg, kwhchiu@hku.hk,
tmc877@ha.org.hk, {chcheung,ssee}@nvidia.com

## Abstract

*Automatic radiology report generation using deep learning models has been recently explored and found promising. Neural decoders are commonly used for the report generation, where irrelevant and unfaithful contents are unavoidable. The retrieval-based approach alleviates the limitation by identifying reports which are relevant to the input to assist the generation. To achieve clinically accurate report retrieval, we make reference to clinicians' diagnostic steps of examining a radiology image where anatomical and diagnostic details are typically focused, and propose a novel hierarchical visual concept representation called anatomy-aware hierarchical vision encoding (AHIVE). To learn AHIVE, we first derive a methodology to extract hierarchical diagnostic descriptions from radiology reports and develop a CLIP-based framework for the model training. Also, the hierarchical architecture of AHIVE is designed to support interactive report retrieval so that report revision made at one layer can be propagated to the subsequent ones to trigger other necessary revisions. We conduct extensive experiments and show that AHIVE can outperform the SOTA vision-language retrieval methods in terms of clinical accuracy by a large margin. We provide also a case study to illustrate how it enables interactive report retrieval.*

## 1. Introduction

Automatic generation of radiology reports aims to assist radiologists with the time-consuming reporting task and expedite the diagnosis workflow. Given a radiology image, the task is to analyze its visual features, identify abnormalities, and generate a diagnostic report accordingly. Clinical accuracy of the generated report is one major goal to achieve. Enabling users to make interactive revision effectively and confirming the report is always desirable.

Deep learning methods have been explored for report generation. One common approach is to adopt the encoder-decoder structure with a convolutional neural network as the encoder and a language model as the decoder [3, 19, 21]. Medical knowledge represented as knowledge graphs [16, 20, 44] and the use of pre-defined clinical templates [14, 34] have been explored to enhance the clinical accuracy. Yet it is unavoidable for the neural decoder to generate irrelevant or unfaithful content [26]. To alleviate the limitation, the retrieval-based approach has been recently explored to retrieve the reports relevant to the input. The retrieved reports can then be used as drafts for further revision to produce the final report [8, 9, 33]. Clinical accuracy of the retrieved reports remains to be one of the key areas to be addressed.

In this paper, inspired by the practice of radiologists in preparing diagnostic reports, we propose a deep retrieval model to encode a radiology image (e.g., X-ray) using a hierarchical visual concept embedding which can capture *anatomical* and *diagnostic* details for *clinically* accurate report retrieval, as illustrated in Fig. 1.
i) *Anatomical details*: Chest-related abnormalities may appear in multiple anatomical parts. Radiologists usually assess all visible anatomical parts to ensure completeness and consistency [4]. While various approaches have been proposed to extract features from different anatomical parts for the report generation [5, 28, 30, 36], the spatial relationship among them is seldom considered.
ii) *Diagnostic details*: Radiologists typically go through steps of examining diagnostic-related visual clues (i.e.,

*Work done during a visiting student at CFAR, Agency for Science, Technology and Research, Singapore.

pathology) to make conclusion for ensuring the report's explainability. Some attempts try to detect and encode the abnormalities, and generate their corresponding sentences for explaining them [16, 22]. Yet, the diagnostic steps of examining an image are often not taken into consideration.

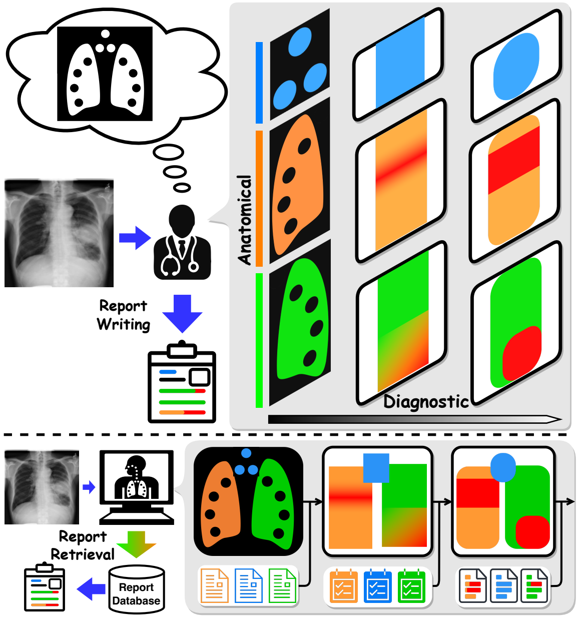


Figure 1. An illustration of how a clinician puts attention on anatomical and diagnostic details for examining a radiology image (upper), and the proposed report retrieval approach (lower).

Furthermore, enabling users to interactively intervene and fine-tune the report retrieval process to improve the retrieval relevancy is desirable [31]. Some existing methods allow users to customize the image regions detected and provide medical keywords [30, 34]. Yet, how to achieve more effective interactive report retrieval intervention is worth further exploration.

To this end, we propose a novel visual representation called anatomy-aware hierarchical vision encoding (AHIVE) to achieve clinically accurate and interactive radiology report retrieval. We first derive for each report a hierarchical diagnostic description with progressive levels of details using some publicly available anatomy-centred annotations (e.g., Chest ImaGemone [38, 39]). The hierarchical description is then used to supervise the learning of AHIVE using the CLIP-based vision-language model framework. For interactive retrieval, the hierarchical design of AHIVE allows the retrieval processes at different levels to be interacting so that revision made by the user at a certain diagnostic level can be propagated to the subsequent levels to trigger further necessary revisions for further improving the retrieval relevancy.

We evaluate the proposed AHIVE-based approach using MIMIC-CXR which is currently the largest publicly available chest x-ray dataset and show that it can outperform the SOTA methods for radiology report retrieval in terms of clinical accuracy by a large margin. The main contributions of this paper are summarised as follows:

- An anatomy-aware hierarchical vision encoding model is proposed for radiology report retrieval, which can be trained based on hierarchical description of radiology reports;
- The proposed model can support interactive report retrieval where the user can just fine-tune the diagnostic details of some retrieval results; and
- We empirically show that with a particular three-level diagnostic description integrated with the CLIP framework, the proposed model outperforms the SOTA CLIP-based retrieval methods in terms of clinical accuracy.

## 2. Related Work

In the literature, various methods based on the encoder-decoder deep architecture have been explored for radiology report generation [3, 19, 21]. To improve clinical accuracy by leveraging prior knowledge, the retrieval-based report generation approach retrieves the clinical templates in the form of medical keywords [29, 42] and diagnostic sentences [1, 15, 24, 35, 41] to assist the generation. As the generation of irrelevant contents by neural decoders is unavoidable, the retrieval-based methods which retrieve the complete reports relevant to the input image have been explored [8, 26, 37]. A specific focus is placed on learning joint multi-modal embedding by multi-granularity visual feature [9, 11, 33]. However, the anatomical and diagnostic features in the visual input are not explicitly considered in the retrieval model. The clinical accuracy of the retrieval results is hard to be ensured. Regarding interactive retrieval of radiology reports, there only exist a few attempts where the medical topics [34] or image regions [30] of the report are allowed to be customized. The flexibility of user intervention to fine-tune retrieval results is still limited.

## 3. Overview of the Proposed Framework

In this paper, we consider a radiology report retrieval system that extracts visual features from a frontal chest X-ray image to form the query to retrieve relevant radiology reports. To achieve clinically accurate report retrieval, we first propose to represent each report as a hierarchical diagnostic description $\{H_0, H_1, ...H_M\}$ where each level is associated with a specific set of diagnostic items. The hierarchy of description starts with diagnostic items about the overall information of different anatomical parts in a radiology image. The level of diagnostic details increases as the level of the hierarchy increases, and $H_M$ is the free-text report (Section 4). We then introduce a novel anatomy-aware hierarchical vision encoding (AHIVE) with a multi-layer architecture that maps the input image $I$ to a hierarchical visual diagnostic embedding $\{Z_m\}_{m=1}^{M}$. It can be learned using the CLIP framework by aligning with the hierarchical diagnostic description $\{H_m\}_{m=0}^{M}$ (Section 5). The hierarchical visual diagnostic embedding $\{Z_m\}$ is then used to retrieve the relevant diagnostic descriptions and then the relevant radiology report.

## 4. Extracting Hierarchical Diagnostic Description from Radiology Report

Radiology diagnosis typically follows some diagnostic steps in examining a radiology image with a focus on different levels of anatomical and diagnostic details in different steps. This motivates our adoption of the hierarchical diagnostic description $\{H_0, H_1, ...H_M\}$. In particular, the diagnostic steps of examining the image lead to the hierarchical representation, and the progressive level of details to be focused is reflected by assigning each description level with a number of associated diagnostic items (e.g., anatomical parts or abnormalities).

Therefore, we represent the diagnostic description at level $m$ as $H_m \in \mathbb{R}^{\mathcal{K}_m \times L}$ which consists of $\mathcal{K}_m$ sentences with at most $L$ tokens each. Each sentence corresponds to one distinct diagnostic item, and thus the level $m$ is associated with $\mathcal{K}_m$ diagnostic items.

Anatomy awareness is important for radiology diagnosis. For instance, the location and size of different anatomical parts often provide clues for diagnosis. We set the description at the level 0 $H_0 \in \mathbb{R}^{\mathcal{K}_0 \times L}$ to be associated with $\mathcal{K}_0 = N$ anatomical parts which are of interest. With that, we can trace back and see how $H_m$ at level $m$ is grounded with different anatomical regions of the image by backward tracking $H_m \rightarrow H_{m-1} ... \rightarrow H_1 \rightarrow H_0$.

While the proposed representation is generic, we adopt a specific representation to be studied in this paper:

$H_0 \in \mathbb{R}^{N \times L}$ contains $N$ diagnostic items which describe whether the size and location of the $N$ anatomical parts are normal or not;
$H_1 \in \mathbb{R}^{N \times L}$ contains $N$ diagnostic items which describe whether the $N$ anatomical part are normal and with some medical devices or not;
$H_2 \in \mathbb{R}^{NK \times L}$ contains $NK$ diagnostic items indicating if a pre-defined set of $K$ abnormalities are detected in $N$ anatomical parts or not; and
$H_3 \in \mathbb{R}^{\mathcal{R} \times L}$ represents the free-text report.

To construct $H_0$, $H_1$ and $H_2$, we make use of the annotations provided by Chest ImaGenome [38, 39] containing the scene graph descriptions of the MIMIC-CXR images with 29 anatomical parts involved. A few templates are manually curated based on anatomy-centered annotations to construct the hierarchical diagnostic description for each report. More implementation details are provided in Section 6.

## 5. Learning AHIVE using Vision-Language Model

Given the hierarchical diagnostic descriptions extracted from the radiology reports in the training set, we propose a novel anatomy-aware hierarchical vision embedding (AHIVE) which can be learned using the vision-language model. As shown in Fig. 2, the model architecture has three components: i) an anatomy-aware vision encoder to compute the spatial embedding of anatomical parts $V_0$, ii) a hierarchical diagnostic vision encoder to compute $M$ vision embeddings $\{Z_m\}_{m=1}^{M}$ correspond to $M$ different levels of diagnostic details, and iii) an $M$-layer vision-text matching module to retrieve radiology reports and their diagnostic descriptions $\{\hat{H}_m\}_{M}^{m=1}$ based on $\{Z_m\}_{m=1}^{M}$.

### 5.1. Anatomy-aware vision encoder

We represent an image $I \in \mathbb{R}^{\mathcal{HW} \times \mathcal{D}}$ as an $\mathcal{H} \times \mathcal{W}$ feature map with $\mathcal{D}$-dimensional features. To achieve anatomy-awareness, we first detect a pre-defined number of anatomical parts using a pre-trained object detector [38, 39] and represent each anatomical part with its own vision embedding. In particular, we first obtain a masking map, denoted as $\mathcal{M}_n^{(\text{Ana})} \in \mathbb{R}^{\mathcal{HW}}$ with $\{0, 1\}$ value indicating whether the bounding box of the detected anatomical part falls on the corresponding feature region.

Instead of modeling the anatomical parts independently, we consider their spatial relationship with reference to the chest area. Taking the chest area as the reference is to avoid the influence of irrelevant features. The masking map of the overall chest area $\mathcal{M}^{(\text{Chest})}$ can be constructed by merging the anatomical masking maps to cover all detected anatomical parts. The vision embeddings of the anatomical part $\{V_n^{(\text{Ana})}\}$ and that of the chest area $V^{(\text{Chest})}$ are computed using spatial attention, given as:

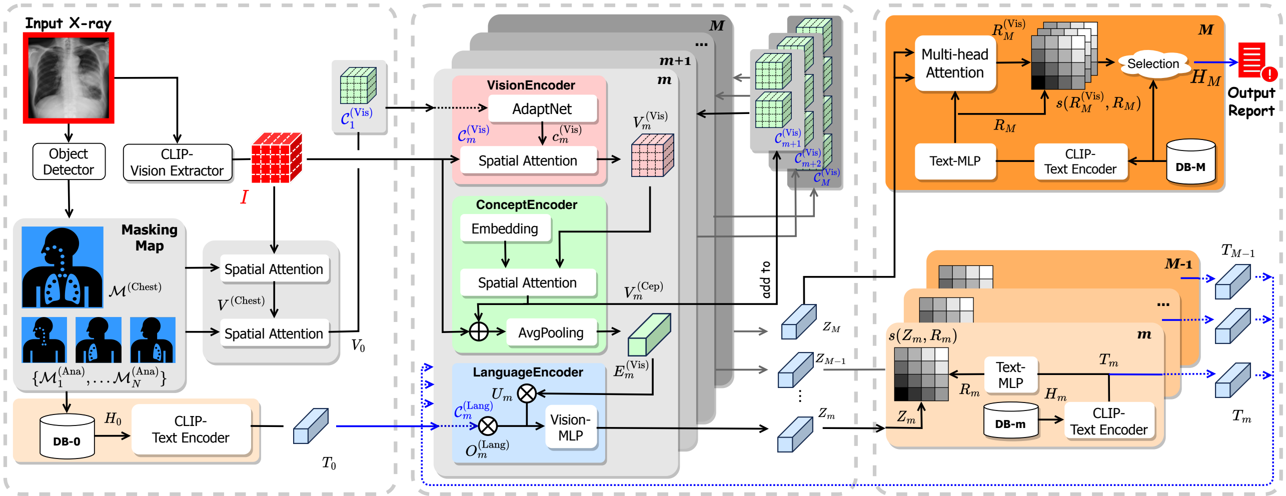
$$\begin{aligned} V^{(\text{Chest})} &= \text{SpaAttn}(I, \mathcal{M}^{(\text{Chest})} \otimes I); \\ V_n^{(\text{Ana})} &= \text{SpaAttn}(V^{(\text{Chest})}, \mathcal{M}_n^{(\text{Ana})} \otimes I), \end{aligned} \tag{1}$$

where $\otimes$ is the operation that applies the same mask $\mathcal{M}$ to all the features of $I$. $\text{SpaAttn}(\cdot, \cdot)$ is the spatial attention network used in [27] which is defined as:

$$\text{SpaAttn}(\mathcal{Q}, \mathcal{V}) = \text{CrossAttn}(\text{STN}(\mathcal{Q}), \mathcal{V}), \tag{2}$$

where $\text{STN}(\cdot)$ is a spatial transformer layer [10] followed by a self-attention layer, $\text{CrossAttn}(\cdot, \cdot)$ is a cross-attention layer followed by a two-layer feed-forward network, the skip connection and the layer normalization.

The overall anatomy-aware vision embedding is obtained by concatenating the vision embeddings of the anatomical parts, given as: $V_0 = \bigoplus_{n=1}^{N} V_n^{(\text{Ana})}$.



i) Anatomy-aware Vision Encoder

ii) Hierarchical Diagnostic Vision Encoder

iii) Descriptions and Report Retrieval

Figure 2. An overall model architecture of the proposed AHIVE for retrieving radiology reports given an X-ray image.

## 5.2. Hierarchical diagnostic vision encoder

Diagnostic radiology reporting typically involves interdependent steps with different diagnostic details to be examined. For instance, questionable image regions are first identified before some specific abnormal observations are further examined. With the motivation that a radiology report is prepared to document the diagnosis process, we propose an $M$-layer hierarchical vision encoder with the incorporation of a hierarchical diagnostic embedding $\{Z_m\}_{m=1}^{M}$ where $Z_m$ refers to the diagnostic embedding of layer $m$.

Specifically, we use the hierarchical description of the diagnostic report to obtain the hierarchical language embedding (Section 4). Then, we learn a hierarchical vision embedding using the CLIP framework. We adopt a vision encoder to compute the hierarchical vision embedding where the dependency among the layers is considered. For each layer, we obtain the vision embedding, project it first to the visual concept space, and then to the visual diagnostic space to facilitate the retrieval. This design allows the hierarchical vision embedding to be learned by effectively aligning with a hierarchical diagnostic description.

Let $V_m^{(\text{Vis})}$, $E_m^{(\text{Vis})}$ and $Z_m$ denote the vision embedding, visual concept embedding and visual diagnostic embedding at layer $m$ respectively, which are computed by

$$\begin{aligned} V_m^{(\text{Vis})} &= \text{VisionEncoder}_m(I, \mathcal{C}_m^{(\text{Vis})}); \\ E_m^{(\text{Vis})} &= \text{ConceptEncoder}_m(V_m^{(\text{Vis})}, E_m); \\ Z_m &= \text{LanguageEncoder}_m(E_m^{(\text{Vis})}, \mathcal{C}_m^{(\text{Lang})}), \end{aligned} \tag{3}$$

where $E_m$ are the learnable embeddings of the medical concepts associated with layer $m$. $\mathcal{C}_m^{(\text{Vis})}$ and $\mathcal{C}_m^{(\text{Lang})}$ are the visual context and language context computed by aggregating the vision and language embeddings in the preceding layers, respectively. The implementation details of the three encoders are presented as follows:

**VisionEncoder$_\mathbf{m}$** computes the *vision embedding* at layer $m$, conditioned to the visual context captured in the preceding layers.

*Implementation*: The visual context, denoted as $\mathcal{C}_m^{(\text{Vis})} = \{V_i^{(\text{Cep})}\}_{i=0}^{m-1}$, contains $m$ concept-aligned vision embeddings which are obtained by ConceptEncoder$_\text{m}$ (to be detailed in the next section).[1] It is first fed to an Adaptive Network [19] (AdaptNet) aggregating the input elements and projects it to an intermediate context embedding $c_m^{(\text{Vis})}$,

$$\text{AdaptNet}(\{V_i^{(\text{Cep})}\}_{i=0}^{m-1}) = \sum_{i=0}^{m-1} \lambda_{m,i}(O_{m,i}^{(\text{Vis})} V_i^{(\text{Cep})}), \tag{4}$$

where $O_{m,i}^{\text{Vis}}$ is a $\mathcal{K}_m \times \mathcal{K}_i$ projection matrix to be learned, and $\lambda_{m,i} \in (0, 1)$ corresponds to the importance weighting of $V_i^{(\text{Cep})}$ for layer $m$. $\lambda_m \in \mathbb{R}^{1 \times m}$ is computed by

$$\lambda_m = \text{sigmoid}(Q_m \sum_{i=0}^{m-1} \delta(O_{m,i}^{(\text{Vis})} V_i^{(\text{Cep})}) W_{m,i}), \tag{5}$$

where $Q_m$ is of dimension $1 \times \mathcal{K}_m \mathcal{HW}$, $\delta(O_{m,i}^{(\text{Vis})} V_i^{(\text{Cep})})$ reshapes the input from $\mathcal{K}_m \times \mathcal{HW} \times \mathcal{D}$ to $\mathcal{K}_m \mathcal{HW} \times \mathcal{D}$, and $W_{m,i}$ is of dimension $\mathcal{D} \times m$. $Q_m$ and $W_{m,i}$ are the learnable parameters and applied together to map $\mathcal{K}_m \mathcal{HW}$ rows to 1 and $\mathcal{D}$ dimension to $m$ elements, respectively.

The vision embedding at layer $m$ $V_m^{(\text{Vis})} \in \mathbb{R}^{\mathcal{K}_m \times \mathcal{HW} \times \mathcal{D}}$ is then obtained by performing spatial attention of $I$ and the context embedding $c_m^{(\text{Vis})}$. Specifically, the vision embedding per each diagnostic item is computed as

$$V_{m|x}^{(\text{Vis})} = \text{SpaAttn}(I, c_{m|x}^{(\text{Vis})}), \tag{6}$$

[1] The visual context of layer 0 $V_0^{(\text{Cep})}$ is set as $V_0$ .

where $x \in [1, \mathcal{K}_m]$ is the index of diagnostic item.

**ConceptEncoder$_\mathbf{m}$** aligns the vision embedding with the medical concepts at layer $m$ to obtain the corresponding *visual concept embedding*.
*Implementation*: We first introduce the medical concept embedding $E_m \in \mathbb{R}^{\mathcal{K}_m \times \mathcal{E} \times \mathcal{D}}$ (to be learned) to encode the $\mathcal{K}_m$ diagnostic items, with $\mathcal{E}$ embedding slots used per diagnostic item. A spatial attention between the vision embedding $V_m^{(\text{Vis})}$ and $E_m$ is then performed to obtain the concept-aligned vision embedding $V_m^{(\text{Cep})}$ to be fed to VisualEncoder$_{m+1}$, given as

$$V_{m|x}^{(\text{Cep})} = \text{SpaAttn}(V_{m|x}^{(\text{Vis})}, E_{m|x}). \tag{7}$$

The visual concept embedding $E_m^{(\text{Vis})} \in \mathbb{R}^{\mathcal{K}_m \times \mathcal{D}}$ is obtained by aggregating over the $\mathcal{H} \times \mathcal{W}$ feature map using global average pooling, given as

$$E_{m|x}^{(\text{Vis})} = \text{FCN}(\underset{\mathcal{HW} \to 1}{\text{AvgPooling}}(V_{m|x}^{(\text{Cep})} \oplus I)), \tag{8}$$

where $V_{m|x}^{(\text{Cep})} \oplus I$ denotes the concatenation on $\mathcal{D}$ dimensions. FCN is a single layer full-connected NN projecting the resulting $2\mathcal{D}$-dimension back to $\mathcal{D}$-dimension.

**LanguageEncoder$_\mathbf{m}$** gives the *visual diagnostic embedding* at layer $m$, conditioned to the current visual context and the language context of retrieved description at layer $m-1$.
*Implementation*: We define the language context $\mathcal{C}_m^{(\text{Lang})} = T_{m-1} \in \mathbb{R}^{\mathcal{K}_{m-1} \times d}$ as the $d$-dimensional text embedding of the description $H_{m-1}$ retrieved at layer $m-1$. Then, we compute a language-enhanced visual diagnostic embedding $E_m^{(\text{Lang})} \in \mathbb{R}^{\mathcal{K}_{m-1} \times \mathcal{D}}$, given as:

$$E_m^{(\text{Lang})} = \text{FFN}(E_m^{(\text{Vis})} U_m + O_m^{(\text{Lang})} T_{m-1}), \tag{9}$$

where $\text{FFN}(\cdot)$ is a two-layer feed-forward network projecting the input back to a $\mathcal{D}$-dimension output, $U_m$ and $O_m^{(\text{Lang})}$ are the learnable parameters. $U_m \in \mathbb{R}^{\mathcal{D} \times d}$ projects $E_m^{(\text{Cep})}$ to $d$-dimension while $O_m^{(\text{Lang})} \in \mathbb{R}^{\mathcal{K}_m \times \mathcal{K}_{m-1}}$ learns the correlation of the diagnostic items between layers $m$ and $m-1$.

We then project $E_m^{(\text{Lang})}$ to a space that aligns with that of the diagnostic descriptions $H_m$ to facilitate matching for subsequent retrieval. With a single layer MLP network adopted for the projection, the final visual diagnostic embedding $Z_m \in \mathbb{R}^{\mathcal{K}_m \times D}$ becomes:

$$Z_m = \text{MLP}_{(\text{Vision})}(E_m^{(\text{Lang})}). \tag{10}$$

### 5.3. Retrieving radiology reports using AHIVE

Given a radiology image, its hierarchical visual diagnostic embedding $\{Z_m\}_{m=1}^M$ can be extracted by AHIVE as the visual query to retrieve sequentially the relevant diagnostic descriptions $\hat{H}_1, \hat{H}_2, ...\hat{H}_{M-1}$ and then radiology report $\hat{H}_M$. Similar to other image-to-text retrieval approaches [8, 25], the relevant description $\hat{H}_m$ is retrieved by measuring the similarity between the visual query $Z_m$ and the textual response $R_m$ obtained by

$$R_m = \text{MLP}_{(\text{Text})}(T_m), \tag{11}$$

where $\text{MLP}_{(\text{Text})}$ is a single layer MLP network project $T_m$ to the $\mathcal{D}$-dimension as $R_m \in \mathbb{R}^{\mathcal{K}_m \times D}$.

**Retrieving hierarchical diagnostic descriptions** Similar to the CLIP architecture [25], the retrieval of the relevant diagnostic description at layer $m$ can be performed by measuring the dot-product similarity $sim(Z, R)$ between the visual diagnostic embedding $Z_m$ and the textual response embedding $R_m$, defined as:

$$sim(Z_m, R_m) = \frac{1}{\mathcal{K}_m} \sum_{x=1}^{\mathcal{K}_m} Z_{m|x} {R_{m|x}}^\top. \tag{12}$$

**Retrieving free-text radiology reports** The diagnostic description $H_M$ at the final layer $M$ corresponds to the free-text radiology report. However, different from the hierarchical descriptions $H_{m<M}$ which are designed to describe vision-relevant diagnostic items, the free-text report unavoidably contains also contents which are not directly corresponding to the visual contents [2, 26]. For example, it may contain elaboration of the findings by comparing with the prior study. It can also be actionable information like recommended follow-up actions.

Therefore, instead of measuring the similarity between $Z_M$ and $R_M$ directly, we instead use multi-head attention to extract vision-related information $R_M^{(\text{Vis})}$ from the report $R_M$ with reference to $Z_M$ and then measure the similarity between $R_M^{(\text{Vis})}$ and $R_M$. Specifically, $R_M^{(\text{Vis})} \in \mathbb{R}^{\mathcal{R} \times D}$ of a given report with $\mathcal{R}$ sentences based on a multi-head attention layer (MHA) can be obtained as:

$$R_M^{(\text{Vis})} = \text{FFN}(\text{MHA}(R_M, Z_M, Z_M)). \tag{13}$$

The matching similarity can then be measured by computing the proportion of vision-related contents retained compared to the overall contents of the report, given as

$$sim(R_M^{(\text{Vis})}, R_M) = \frac{1}{\mathcal{R} \times \mathcal{R}} \sum_{i=1}^{\mathcal{R}} \sum_{j=1}^{\mathcal{R}} R_{M|i}^{(\text{Vis})} {R_{M|j}}^\top. \tag{14}$$

**Retrieval with user interaction** The proposed AHIVE is deliberately designed with the diagnostic vision embedding $Z_m$ conditioned to both the visual and language contexts which can be intervened in real time by the user via modifying the hierarchical descriptions $\{H_m\}$.

Accordingly, by modifying the retrieved descriptions up to layer $m-1$ which are decoded as the language context, different interventions to the retrieval of $H_m$ can be carried out. For instance, one could modify the retrieved description at one of the layers, that is $H_i \in \{H_i\}_{i=0}^{m-1}$, for any associated diagnostic items, which will then trigger revised retrieval of $H_{i+1}$ to $H_m$. Also, one could modify the retrieved descriptions of some specific diagnostic items (e.g., abnormalties) $\{H_{i|x}\}_{x=1,i<m}^{\mathcal{K}_i}$ up to layer $m-1$ to support more focused intervention by the user.

As the aforementioned modifications can be simply performed by editing the retrieved diagnostic descriptions in nature language, it allows users (e.g., radiologists in the hospital) to easily interact via a simple user interface to refine the retrieval process.

# 6. Experiments

## 6.1. Data and evaluation metrics

We use the largest publicly available radiology report dataset MIMIC CXR dataset [12, 13] for performance evaluation. We extract findings/impression sections as the target report and tokenize them with the maximum length set as 220 covering the whole reports of 99.9% samples. Following the original split setting of the dataset, the training/validation/test size is split to 222,705 / 1,807 / 3,269.

To evaluate the clinical quality of the reports retrieved, we adopt the clinical efficacy metrics (CE) [3] and the radiology report quality index (RadRQI-F1) [40]. i) For CE, as in [3, 15], we report the micro-average F1 scores of `CE(11)`, `CE(11/5)`, and `CE(13+NL)` with the suffix `([# Labels])` indicating the number of abnormalities being evaluated. ii) For RadRQI-F1, we measure the scores based on the same abnormalities adopted in `CE(11)` as well as the top-50 abnormalities. RadRQI-F1(Hits) refers to the average number of classes which have non-zero F1 scores. We also adopt the common NLP metrics like BLEU [23], ROUGE [17] and CIDEr [32] to evaluate language quality of the retrieved reports. For all the metrics, we report the average performance scores of three runs with different random seeds in all our experiments.

## 6.2. Experiment settings and baselines

To evaluate the performance of the proposed AHIVE for radiology report retrieval, we construct the hierarchical diagnostic descriptions for all the radiology reports in the dataset and learn the proposed AHIVE as presented in Sections 4 and 5. We compare its retrieval performance against several image-text retrieval methods as the baselines.

**Experiment Settings**: We construct the three-level diagnostic description $\{H_m\}_{m=0}^{M=2}$ for each report as described in Section 4. Fig. 3 shows an example. We focus on the five anatomical parts as considered in `CE(11/5)`. The numbers of diagnostic items $\mathcal{K}_0$, $\mathcal{K}_1$ and $\mathcal{K}_2$ are set to be 5, 5 and 35[2]. The maximum number of tokens $L$ is set to 20.

| Layer | Diagnosis | Diagnostic Description |
|---|---|---|
| 0 | *Anatomical Property* | "Left lung is out of the normal location.<br>Left lung is out of the normal size." |
| 1 | *Anatomical Health* | "Left lung is in abnormal.<br>There is no medical device in left lung." |
| 2 | *Anatomical Abnormality* | "There is no airspace opacity in left lung."<br>"There is no pneumothorax in left lung."<br>"There is no mass/nodule in left lung."<br>"There is no device or tube in left lung."<br>"There is lung opacity in left lung."<br>"There is no lung lesion in left lung."<br>"There is pleural effusion in left lung."<br>"There is no consolidation in left lung."<br>"There is pulmonary edema in left lung."<br>"There is no atelectasis in left lung." |

Figure 3. Illustration of a hierarchical diagnostic description regarding the anatomical part *Left Lung*. The diagnostic items for each level are underlined and the negation indicator is in red color.

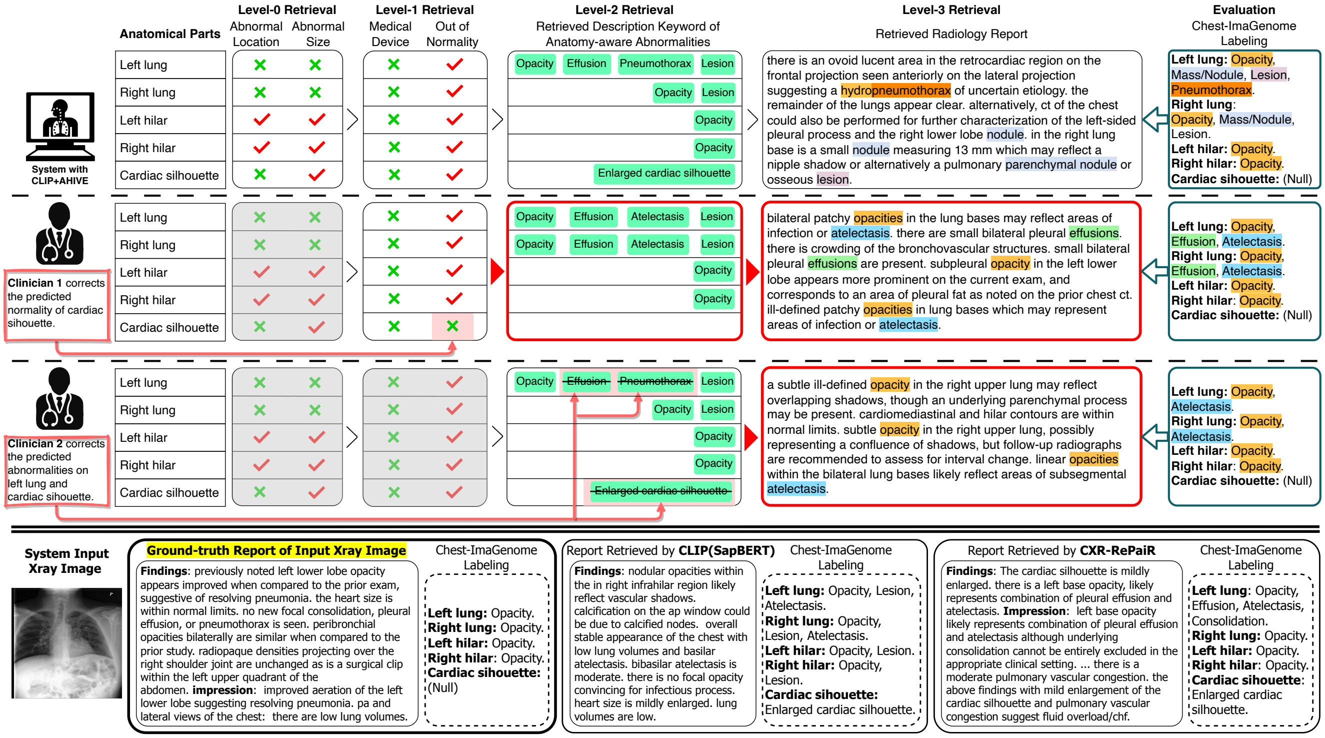
For the base model CLIP [25] which is used for learning AHIVE, we adopt the pre-trained SapBERT [18] and ViT [7] as the text and image encoder backbones, respectively. We denote this base model as `CLIP(SapBERT)`. In our experiments, all input images are resized to 224 × 224 before feeding into the ViT. For memory efficiency, we adopt two-phrase training to learn `CLIP(SapBERT)+AHIVE`. We first finetune `CLIP(SapBERT)` based on MIMIC CXR dataset and then learn `AHIVE` using the finetuned-and-frozen `CLIP(SapBERT)`. The number of epochs and batch size are set to 20 and 512, respectively. The optimizer is AdamW with learning rate of $1e$-6. More implementation details could be found in supplementary materials.

**Baselines**: To evaluate the performance achieved by integrating AHIVE with CLIP, we compare with several pre-trained CLIPs, including: i) `CLIP(Vanilla)` which uses ViT and vanilla transformer as vision and text encoders pre-trained on Imagenet [6]; ii) `CLIP(SapBERT)` which is fully fine-tuned on MIMIC CXR; and iii) `CLIP(SapBERT/Frozen)` which has the same encoders with `CLIP(SapBERT)` while fine-tuning only its projection layers on MIMIC CXR and keeping the rest frozen. We also test several state-of-the-art approaches, including: `CXR-RePaiR` (2021) [8], `MedCLIP` (2022) [37], `BiomedCLIP` (2023) [43] and `X-REM` (2023) [11] (see the supplementary materials for details).

## 6.3. Performance on report retrieval

Table 1 shows the performance comparison results between the CLIP model integrated with `AHIVE` and the other CLIP

[2] $\mathcal{K}_2$ is set as 35 as there are 35 distinct pairs of (anatomical part, abnormality) for the 11 abnormalities and 5 anatomical parts considered in `CE(11/5)`.



there is an ovoid lucent area in the retrocardiac region on the frontal projection seen anteriorly on the lateral projection suggesting a hydropneumothorax of uncertain etiology. the remainder of the lungs appear clear. alternatively, ct of the chest could also be performed for further characterization of the left-sided pleural process and the right lower lobe nodule. in the right lung base is a small nodule measuring 13 mm which may reflect a nipple shadow or alternatively a pulmonary parenchymal nodule or osseous lesion.

**Left lung:** Opacity, Mass/Nodule, Lesion, Pneumothorax.
**Right lung**: Opacity, Mass/Nodule, Lesion.
**Left hilar:** Opacity.
**Right hilar:** Opacity.
**Cardiac sihouette:** (Null)

bilateral patchy opacities in the lung bases may reflect areas of infection or atelectasis. there are small bilateral pleural effusions. there is crowding of the bronchovascular structures. small bilateral pleural effusions are present. subpleural opacity in the left lower lobe appears more prominent on the current exam, and corresponds to an area of pleural fat as noted on the prior chest ct. ill-defined patchy opacities in lung bases which may represent areas of infection or atelectasis.

**Left lung:** Opacity, Effusion, Atelectasis.
**Right lung:** Opacity, Effusion, Atelectasis.
**Left hilar:** Opacity.
**Right hilar**: Opacity.
**Cardiac sihouette:** (Null)

a subtle ill-defined opacity in the right upper lung may reflect overlapping shadows, though an underlying parenchymal process may be present. cardiomediastinal and hilar contours are within normal limits. subtle opacity in the right upper lung, possibly representing a confluence of shadows, but follow-up radiographs are recommended to assess for interval change. linear opacities within the bilateral lung bases likely reflect areas of subsegmental atelectasis.

**Left lung:** Opacity, Atelectasis.
**Right lung:** Opacity, Atelectasis.
**Left hilar:** Opacity.
**Right hilar**: Opacity.
**Cardiac sihouette:** (Null)

**Ground-truth Report of Input Xray Image**

**Findings**: previously noted left lower lobe opacity appears improved when compared to the prior exam, suggestive of resolving pneumonia. the heart size is within normal limits. no new focal consolidation, pleural effusion, or pneumothorax is seen. peribronchial opacities bilaterally are similar when compared to the prior study. radiopaque densities projecting over the right shoulder joint are unchanged as is a surgical clip within the left upper quadrant of the abdomen. **impression**: improved aeration of the left lower lobe suggesting resolving pneumonia. pa and lateral views of the chest: there are low lung volumes.

Chest-ImaGenome Labeling

**Left lung:** Opacity.
**Right lung:** Opacity.
**Left hilar:** Opacity.
**Right hilar**: Opacity.
**Cardiac sihouette:** (Null)

Report Retrieved by **CLIP(SapBERT)**

**Findings**: nodular opacities within the in right infrahilar region likely reflect vascular shadows. calcification on the ap window could be due to calcified nodes. overall stable appearance of the chest with low lung volumes and basilar atelectasis. bibasilar atelectasis is moderate. there is no focal opacity convincing for infectious process. heart size is mildly enlarged. lung volumes are low.

Chest-ImaGenome Labeling

**Left lung:** Opacity, Lesion, Atelectasis.
**Right lung:** Opacity, Lesion, Atelectasis.
**Left hilar:** Opacity, Lesion.
**Right hilar:** Opacity, Lesion.
**Cardiac sihouette:** Enlarged cardiac silhouette.

Report Retrieved by **CXR-RePaiR**

**Findings**: The cardiac silhouette is mildly enlarged. there is a left base opacity, likely represents combination of pleural effusion and atelectasis. **Impression**: left base opacity likely represents combination of pleural effusion and atelectasis although underlying consolidation cannot be entirely excluded in the appropriate clinical setting. ... there is a moderate pulmonary vascular congestion. the above findings with mild enlargement of the cardiac silhouette and pulmonary vascular congestion suggest fluid overload/chf.

Chest-ImaGenome Labeling

**Left lung**: Opacity, Effusion, Atelectasis, Consolidation.
**Right lung:** Opacity.
**Left hilar:** Opacity.
**Right hilar**: Opacity.
**Cardiac sihouette**: Enlarged cardiac silhouette.

Figure 4. Illustration of the advantages of AHIVE over the baselines. Rows 1-3: Three retrieval processes with modifications by user via AHIVE. Row 4: The input X-ray image, the ground-truth report and the reports retrieved by two baselines. The unchanged and modified items are colored gray and red, respectively. The updated retrieval results are circled by the red rounded rectangle.

| Model | Clinical Efficacy (# labels) | | | RadRQI-F1 (# labels) | | | NLG | | |
|---|---|---|---|---|---|---|---|---|---|
| | (11/5) | (11) | (13+NL) | (11) | (Top-50) | Hits | B. | R. | C. |
| CLIP(Vanilla) [25] | 0.361 | 0.407 | 0.554 | 0.131 | 0.191 | 9 | 0.046 | 0.132 | 0.011 |
| CLIP(SapBERT/Frozen) [25] | 0.474 | 0.548 | 0.574 | 0.194 | 0.239 | 28 | 0.099 | 0.143 | 0.093 |
| CLIP(SapBERT) [25] | 0.569 | 0.642 | 0.601 | 0.242 | 0.292 | **43** | **0.202** | **0.217** | <u>0.353</u> |
| BiomedCLIP [43] | 0.455 | 0.545 | 0.627 | 0.284 | 0.260 | 21 | 0.077 | 0.153 | 0.065 |
| MedCLIP [37] | 0.534 | 0.618 | 0.596 | 0.228 | 0.170 | 17 | 0.089 | 0.128 | 0.013 |
| X-REM [11] | 0.481 | 0.538 | 0.617 | 0.343 | 0.297 | 38 | 0.127 | 0.178 | 0.352 |
| CXR-RePaiR [8] | 0.564 | <u>0.654</u> | <u>0.660</u> | <u>0.347</u> | <u>0.309</u> | <u>40</u> | <u>0.160</u> | <u>0.183</u> | **0.359** |
| CLIP(SapBERT)+AHIVE (**ours**) | **0.604** | **0.678** | **0.685** | **0.407** | **0.327** | 30 | 0.131 | 0.163 | 0.324 |

Table 1. Performance comparison on report retrieval based on clinical accuracy and NLG metrics. The best scores are in bold face and the second best are underlined. "B.", "R." and "C." stand for BLEU (average), ROUGE-L and CIDEr scores.

baselines in terms of clinical accuracy using the MIMIC CXR dataset. Among the baselines, CLIP(SapBERT) gives the best or close to the best overall performance based on clinical accuracy metrics CE and RadRQI-F1.

With AHIVE incorporated into CLIP(SapBERT), the best performance is achieved in terms of anatomy-centered accuracy CE(11/5), global accuracy CE(11) and normality-included accuracy CE(13+NL). We also observe performance improvement of 68.18% and 11.99% over CLIP(SapBERT) based on RadRQI-F1(11) and RadRQI-F1(50) which evaluate also the relevancy of the attributes associated with abnormalities in the retrieved reports. This hints that incorporating AHIVE can effectively improve the retrieval quality for both abnormalities and their associated attributes.

It is to be noted that CLIP(SapBERT)+AHIVE is not as good as CLIP(SapBERT) based on the metric RadRQI(Hits). It is probably due to the limited number of abnormalities considered by AHIVE in this experiment. Since the visual diagnostic embedding is encoded given the language context of a specific set of pre-defined abnormalities, the candidate reports with abnormalities not considered by AHIVE could be discounted even though they can be highly relevant to the input X-ray image, which results in the lower `RadRQI(Hits)` score. How to select the pre-defined abnormalities to achieve a high abnormality coverage of the dataset remains open.

| $N$ | $K$ | CE | | RadRQI | |
|---|---|---|---|---|---|
| | | $(K/N)$ | $(K)$ | $(K)$ | (Top-50) |
| 3 | 11 | **0.611*** | 0.663 | 0.384 | 0.318 |
| 5 | 11 | 0.604 | **0.678*** | **0.407** | 0.327 |
| 9 | 11 | 0.528 | 0.655 | 0.401 | **0.331** |
| 5 | 5 | 0.602 | 0.589 | 0.391 | 0.294 |
| 5 | 11 | 0.604 | **0.678*** | 0.407 | 0.327 |
| 5 | 16 | **0.637**** | 0.621 | **0.409** | **0.355*** |

Table 2. Performance comparison with significant test for variants of AHIVE with diagnostic descriptions covering different numbers of anatomical parts ($N$) and abnormalities ($K$).

**Sensitivity analysis** To understand how the performance is affected by different settings of the diagnostic description, we test the proposed `AHIVE` with the diagnostic description covering different numbers of anatomical parts and abnormalities (as shown in Table 2). According to the `RadRQI-F1`, we observe the `AHIVE` incorporating diagnostic descriptions with more anatomical parts and abnormalities in general can achieve higher clinical accuracy. In addition, as shown in `CE(`$K/N$`)`, we notice that higher improvement can be achieved by the diagnostic description with a higher ratio of abnormalities per anatomical part. Yet, we observe some exceptions. How to obtain the optimal setting for the diagnostic description remains open.

**Ablation study** To evaluate the importance of introducing visual and language contexts into `AHIVE`, we test some variants of `AHIVE` without the visual and language contexts at different layers. The results of performance degradation are shown in Table 3. Comparing with the visual context, we observe that `AHIVE` without the language context encounters a higher degree of performance degradation. This observation in turn indicates that the user interactivity provided by `AHIVE` to modify the retrieved diagnostic description is an effective mean of fine-tuning the retrieval results.

**Case study** Fig. 4 illustrates a case study of retrieving reports using AHIVE and two baselines. We present three retrieval processes of the proposed models with: i) no intervention, ii) intervention at $1^{st}$ layer, and iii) interaction at $2^{nd}$ layer. As observed, given an X-ray image of *Opacity*, the relevant reports retrieved by both baselines and AHIVE (with no intervention) cover many irrelevant abnormalities.

| **AHIVE w/o** $\mathcal{C}_m^{(*)}$ | $\downarrow$ **ΔCE**(%) | | $\downarrow$ **ΔRadRQI**(%) | |
|---|---|---|---|---|
| | (11/5) | (11) | (11) | (Top-50) |
| $\mathcal{C}_1^{(\text{Vis})}$ | 0.3 | $\underline{0.3}$ | $\underline{0.5}$ | 0.5 |
| $\mathcal{C}_1^{(\text{Lang})}$ | **0.7** | **0.8** | $\underline{0.5}$ | $\underline{1.0}$ |
| $\mathcal{C}_1^{(\text{Vis})}, \mathcal{C}_1^{(\text{Lang})}$ | $\underline{0.5}$ | $\underline{0.3}$ | **1.4*** | **2.1** |
| $\mathcal{C}_2^{(\text{Vis})}$ | 0.2 | 0.3 | 0.5 | 2.1 |
| $\mathcal{C}_2^{(\text{Lang})}$ | $\underline{4.0}$ | $\underline{3.6}$ | **4.3*** | $\underline{3.7}$ |
| $\mathcal{C}_2^{(\text{Vis})}, \mathcal{C}_2^{(\text{Lang})}$ | **7.3*** | **4.4** | $\underline{2.9}$ | **6.8**** |
| $\mathcal{C}_3^{(\text{Vis})}$ | 0.2 | 0.2 | 0.0 | 1.6 |
| $\mathcal{C}_3^{(\text{Lang})}$ | **4.2** | **3.5** | $\underline{1.9}$ | $\underline{1.6}$ |
| $\mathcal{C}_3^{(\text{Vis})}, \mathcal{C}_3^{(\text{Lang})}$ | $\underline{3.7}$ | $\underline{3.3}$ | **3.8**** | **4.2**** |

Table 3. Performance comparison of variants of AHIVE with or without visual and language contexts introduced at each layer. A larger number indicates a bigger drop in clinical accuracy.

To refine the retrieval results, two cases of modifying descriptions at different layers are shown: i) Intervention at layer $m$=1 (at $2^{nd}$ row) is performed by correcting the normality prediction of *Cardiac Silhouette*. This leads to the diagnostic description at layer $m$=2 to be further updated, and another report with less out-of-target abnormalities covered is then retrieved. ii) Intervention at layer $m$=2 (at $3^{rd}$ row) is performed by correcting three error abnormalities predicted on the *Left Lung* and *Cardiac Silhouette*. A report with more precise abnormalities is then retrieved.

## 7. Conclusion

We propose an anatomy-aware hierarchical vision encoding called AHIVE which can be learned under the CLIP framework using the hierarchical diagnostic descriptions extracted from radiology reports. A particular AHIVE model learned with reference to a three-level diagnostic description outperforms the SOTA CLIP-based retrieval methods in terms of clinical accuracy. It also supports real-time user intervention to fine-tune the retrieval result interactively. AHIVE also possesses Future research possibilities include extending the retrieval approach to the retrieval-based report generation so that the user's fine-tuning can be further reduced. *limitations*: Manual effort is required to design templates for extracting the hierarchical diagnostic information from reports for training, which could be sub-optimal. Also, AHIVE prefers radiology reports with more sentences mentioning the visual clues due to the design of the multi-head attention for the retrieval, which may discount the reports with more elaboration on prior studies and actionable information, even if they are highly relevant.

## References

[1] Siddharth Biswal, Cao Xiao, Lucas Glass, Brandon Westover, and Jimeng Sun. Clara: Clinical report auto-completion. In *Proceedings of the International World Wide Web Conference*, pages 541–550, 2020. 2

[2] Michael A Bruno, Jonelle Petscavage-Thomas, and Hani H Abujudeh. Communicating uncertainty in the radiology report. *American Journal of Roentgenology*, 209(5):1006–1008, 2017. 5

[3] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1439–1449, 2020. 1, 2, 6

[4] Rajat Chowdhury, Iain Wilson, Christopher Rofe, and Graham Lloyd-Jones. *Radiology at a Glance*. John Wiley & Sons, 2017. 1

[5] F. Dalla Serra, C. Wang, F. Deligianni, J. Dalton, and A.Q O'Neil. Finding-aware anatomical tokens for chest x-ray automated reporting. In *MLMI'23*, pages 413–423, 2023. 1

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 6

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[8] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021. 1, 2, 5, 6, 7

[9] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 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, 2021. 1, 2

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in Neural Information Processing systems*, 28, 2015. 3

[11] Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Subathra Adithan, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation. In *Medical Imaging with Deep Learning*, 2023. 2, 6, 7

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

[13] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 6

[14] Ming Kong, Zhengxing Huang, Kun Kuang, Qiang Zhu, and Fei Wu. Transq: Transformer-based semantic query for medical report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 610–620. Springer, 2022. 1

[15] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 1530–1540, 2018. 2, 6

[16] Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *Proceedings of the Conference of Association for the Advance of Artificial Intelligence*, pages 6666–6673, 2019. 1, 2

[17] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004. 6

[18] Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. Self-alignment pretraining for biomedical entity representations. *arXiv preprint arXiv:2010.11784*, 2020. 6

[19] Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 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, 2021. 1, 2, 4

[20] Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, 34, 2021. 1

[21] Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, 2021. 1, 2

[22] Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. *arXiv e-prints*, pages arXiv–2102, 2021. 2

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceeding of the Conference of Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[24] Pablo Pino, Denis Parra, Cecilia Besa, and Claudio Lagos. Clinically correct report generation from chest x-rays using templates. In *International Workshop on Machine Learning in Medical Imaging*, pages 654–663. Springer, 2021. 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5, 6, 7

[26] Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pages 456–473. PMLR, 2022. 1, 2, 5

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceeding of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10674–10685, 2022. 3

[28] Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeffrey Dalton, and Alison Q O'Neil. Controllable chest x-ray report generation from longitudinal representations. *arXiv preprint arXiv:2310.05881*, 2023. 1

[29] Luyao Shi, Tanveer Syeda-mahmood, and Tyler Baldwin. Improving neural models for radiology report retrieval with lexicon-based automated annotation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3457–3463. Association for Computational Linguistics, 2022. 2

[30] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7433–7442, 2023. 1, 2

[31] Phuong Dung Trieu, Sarah J Lewis, Tong Li, Karen Ho, Dennis J Wong, Oanh TM Tran, Louise Puslednik, Deborah Black, and Patrick C Brennan. Improving radiologist's ability in identifying particular abnormal lesions on mammograms through training test set with immediate feedback. *Scientific Reports*, 11(1):9899, 2021. 2

[32] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015. 6

[33] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *arXiv preprint arXiv:2210.06044*, 2022. 1, 2

[34] Lin Wang, Munan Ning, Donghuan Lu, Dong Wei, Yefeng Zheng, and Jie Chen. An inclusive task-aware framework for radiology report generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 568–577. Springer, 2022. 1, 2

[35] Xuwen Wang, Yu Zhang, Zhen Guo, and Jiao Li. Tmrgm: A template-based multi-attention model for x-ray imaging report generation. *Journal of Artificial Intelligence for Medical Sciences*, 2021. 2

[36] Yuhao Wang, Kai Wang, Xiaohong Liu, Tianrun Gao, Jingyue Zhang, and Guangyu Wang. Self adaptive global-local feature enhancement for radiology report generation. In *Proceeding of the IEEE International Conference on Image Processing*, pages 2275–2279. IEEE, 2023. 1

[37] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022. 2, 6, 7

[38] Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021. 2, 3

[39] Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset (version 1.0. 0). *PhysioNet*, 5:18, 2021. 2, 3

[40] Sixing Yan, William K. Cheung, Keith Chiu, Terence M. Tong, Ka Chun Cheung, and Simon See. Attributed abnormality graph embedding for clinically accurate x-ray report generation. *IEEE Transactions on Medical Imaging*, pages 1–1, 2023. 6

[41] Xingyi Yang, Muchao Ye, Quanzeng You, and Fenglong Ma. Writing by memorizing: Hierarchical retrieval-based medical report generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5000–5009. Association for Computational Linguistics, 2021. 2

[42] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 721–729, 2019. 2

[43] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023. 6, 7

[44] Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. In *Proceedings of the Conference of Association for the Advance of Artificial Intelligence*, pages 12910–12917, 2020. 1