RIHA: Report-Image Hierarchical Alignment for Radiology Report Generation

Yucheng Chen, Yang Yu, Yufei Shi, Conghao Xiong, Xulei Yang, and Si Yong Yeo

***Abstract*— Radiology report generation (RRG) has emerged as a promising approach to alleviate radiologists' workload and reduce human errors by automatically generating diagnostic reports from medical images. A key challenge in RRG is achieving fine-grained alignment between complex visual features and the hierarchical structure of long-form radiology reports. Although recent methods have improved image-text representation learning, they often treat reports as flat sequences, overlooking their structured sections and semantic hierarchies. This simplification hinders precise cross-modal alignment and weakens RRG accuracy. To address this challenge, we propose RIHA (Report-Image Hierarchical Alignment Transformer), a novel end-to-end framework that performs multi-level alignment between radiological images and their corresponding reports across paragraph, sentence, and word levels. This hierarchical alignment enables more precise cross-modal mapping, essential for capturing the nuanced semantics embedded in clinical narratives. Specifically, RIHA introduces a Visual Feature Pyramid (VFP) to extract multi-scale visual features and a Text Feature Pyramid (TFP) to represent multi-granularity textual structures. These components are integrated through a Cross-modal Hierarchical Alignment (CHA) module, leveraging optimal transport to effectively align visual and textual features across various levels. Furthermore, we incorporate Relative Positional Encoding (RPE) into the decoder to model spatial and semantic relationships among tokens, enhancing the token-level alignment between visual features and generated text. Extensive experiments on two benchmark chest X-ray datasets, IU-Xray and MIMIC-CXR, demonstrate that RIHA outperforms existing state-of-the-art models in both natural language generation and clinical efficacy metrics.**

***Index Terms*— Radiology report generation, Cross-modal alignment, Visual language models, Radiomics, Precision medical imaging, Multi-granularity alignment**

I. Introduction

RADIOLOGY report generation (RRG) aims to produce accurate free-text descriptions of radiology images (e.g., chest X-rays) to support clinical decision-making and patient treatment. In clinical practice, interpreting radiographic images and composing diagnostic reports is a labor-intensive task that demands the specialized expertise of radiologists. This challenge is further compounded by a global shortage of radiology professionals [1]. An exemplar radiology report with multiple granularities is also shown in Fig. 1. Unlike natural images, radiological scans are characterized by high structural complexity, abstraction, and domain-specific visual patterns, making them more difficult to describe and interpret. As a result, establishing direct and well-defined structural correspondences between visual content and textual descriptions becomes particularly challenging [2]. Thus, automated radiology report generation has emerged as a promising solution to reduce radiologists' workload while preserving the accuracy and quality of the reports [3].

Recent advancements in artificial intelligence have significantly propelled research in RRG [3], [4], offering potential to alleviate radiologists' workloads and address staffing shortages in hospitals [1]. While RRG is often framed similarly to image captioning [5], substantial differences introduce unique challenges when applying conventional techniques to medical contexts [6]. Unlike image captioning that generates balanced descriptions of visual concepts, RRG emphasizes detecting abnormal findings within predominantly normal anatomical structures [7], and involves generating multi-sentence, paragraph-length reports with complex semantic relationships [8]. These extended narratives often contain complex semantic relationships that pose additional challenges for models to capture fine-grained alignments between visual and textual modalities. Crucially, the hierarchical structure of radiology reports directly reflects the clinical diagnostic workflow, where radiologists systematically process images from global anatomical assessment to regional abnormality detection to precise quantitative measurements [9]. This diagnostic approach manifests in reports where paragraph-level impressions provide overall clinical context, sentence-level descriptions detail specific anatomical findings, and word-level terminology captures precise medical measurements. The inherent hierarchical nature of medical reasoning demands sophisticated visual-textual alignment mechanisms that can capture multi-level correspondences essential for clinically coherent report generation.

Existing approaches to address these alignment challenges

Yucheng Chen, Yufei Shi, and Si Yong Yeo are with MedVisAI Lab, Lee Kong Chian School of Medicine, Nanyang Technological University (NTU) ,and Centre of AI in Medicine (email: yucheng005@e.ntu.edu.sg; yufei005@e.ntu.edu.sg; siyong.yeo@ntu.edu.sg).

Yang Yu and Xulei Yang are with Department of Machine Intellection, Institute for Infocomm Research (I^2R), Agency for Science, Technology and Research, (A*STAR), Singapore (email: yu_yang@i2r.a-star.edu.sg; yang_xulei@i2r.a-star.edu.sg).

Conghao Xiong is with the Department of Computer Science and Engineering, the Chinese University of Hong Kong, Hong Kong SAR, China (email: chxiong21@cse.cuhk.edu.hk).

(Si Yong Yeo is the corresponding author).

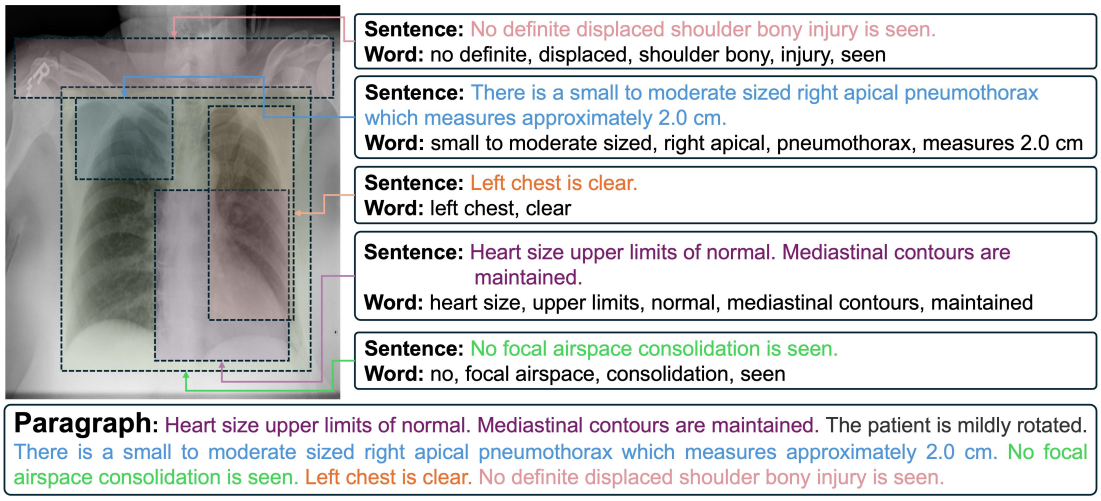

Fig. 1. A chest X-ray image alongside its corresponding radiology report. The boxes on the right display, respectively, the individual sentences and the extracted keywords from the report. Multi-level visual-textual alignments are indicated using matching colours to highlight their associations.

broadly fall into two categories: image-report-only methods [10]–[14] that rely solely on available visual and textual data, and augmented knowledge approaches [15]–[21] that incorporate external medical domain knowledge. In image-report-only methods, current multi-granularity approaches primarily adopt sequential or auxiliary task-based paradigms that fail to achieve comprehensive cross-modal correspondence. Li *et al.* [22] employ a three-stage sequential framework that processes the alignment phase independently. Liu *et al.* [23] rely on sentence-level contrastive learning without extending to comprehensive hierarchical alignment. These methods typically focus on single-modal granularity processing - either visual or textual hierarchies - missing the opportunity to establish synchronized cross-modal correspondences. Moreover, existing hierarchical approaches process different semantic levels sequentially or independently, failing to capture the comprehensive multi-level visual-textual mappings necessary for maintaining clinical coherence. This fundamental gap between existing alignment paradigms and the simultaneous multi-granularity nature of medical reasoning motivates the need for novel approaches that can concurrently establish correspondence across all semantic levels through unified cross-modal hierarchical alignment.

To address the challenges of aligning complex visual and textual information in radiological imaging for the report generation task, we propose a Report-Image Hierarchical Alignment Transformer (RIHA), a novel framework designed to perform hierarchical alignment between radiological images and associated reports at three textual levels: paragraph, sentence, and word. This multi-level alignment enables fine-grained cross-modal mapping, which is essential for capturing the nuanced yet crucial semantics of radiology reports. RIHA also leverages a pyramid-based structure to encode visual data into multi-scale hierarchical representations, thereby facilitating a more accurate and semantically consistent alignment with the multi-granularity textual inputs. This design is especially beneficial for radiological vision-language models, where conventional methods often struggle due to the inherent mismatch between textual semantics and visual structures. The main contributions of this work are summarized as follows:

1) We propose a novel end-to-end Report-Image Hierarchical Alignment (RIHA) method for improving fine-grained alignment between paired X-ray images and long-text radiology reports. RIHA captures hierarchical correspondences at three levels of granularity: shallow (word-level), intermediate (sentence-level), and high (paragraph-level) features.
2) We design a Visual Feature Pyramid (VFP) and a Text Feature Pyramid (TFP) extractor to capture multi-granularity visual and textual features. These modules coordinate with a Cross-modal Hierarchical Alignment (CHA) module, leveraging optimal transport to align semantic information across visual and textual modalities effectively.
3) To enhance token-level alignment between visual features and generated reports, we also integrate a Relative Positional Encoding (RPE) strategy into the decoder. This approach allows the model to generate contextually accurate radiology descriptions that closely correspond to the underlying visual features by capturing both spatial and semantic token relationships, thus improving semantic consistency and spatial accuracy.
4) We conduct comprehensive experiments on two publicly available X-ray datasets to demonstrate the substantial performance enhancements of our method over existing approaches in the report generation task. Additionally, we validate the robustness and adaptability of our approach to practical scenarios through a series of detailed ablation studies.

II. Related Work

A. Image Captioning

Image captioning, as the task of automatically generating natural language descriptions for images, has become a prominent research topic at the intersection of computer vision [24]–[27] and natural language processing [28], [29].

The predominant approach in this field adopts an encoder-decoder framework, in which a visual encoder (typically based on convolutional neural networks or Vision Transformers) first extracts semantic features from the input image [30]. These features are then fed into a language decoder, often implemented using recurrent neural networks or Transformer models, to generate coherent textual descriptions [11], [31]. While early models relied on standard supervised learning, subsequent efforts have introduced more advanced techniques, such as attention mechanisms [32], reinforcement learning [33], and adversarial training [34], to enhance the quality and relevance of generated captions. More recently, the field has seen a shift toward large-scale vision-and-language pretraining (VLP) models, which leverage vast image-text corpora to learn rich multimodal representations, achieving state-of-the-art (SOTA) performance across various benchmarks [35]–[37].

B. Radiology Report Generation

The automatic generation of clinically accurate radiology reports has become a pivotal area of research. Recent approaches to medical report generation can be broadly categorized into two types based on their training data requirements and architectural frameworks: image-report-only and knowledge-augmentation-based paradigms. The knowledge-augmentation-based paradigm enhances model training by incorporating external medical domain knowledge to address the limitations of purely data-driven approaches that may lack essential clinical insight. For instance, Yang *et al.* [20], [21] propose to enhance radiology report generation by incorporating medical knowledge: one learns a medical knowledge base with multi-modal alignment, and the other incorporates both general and specific medical knowledge. [15], [16] all employ the knowledge graph as the external knowledge to help the report generation. Another augmentation strategy is to incorporate external reference reports from existing databases, known as retrieval-augmented methods like [38], [39]. While these methods can improve performance, they introduce substantial practical challenges, including the need for specialized medical annotations and complex pipeline engineering. Our method belongs to the image-report-only paradigm, which relies exclusively on paired image-report data. Approaches within this paradigm treat reports as flat sequences, employing conventional attention mechanisms [40] for visual-textual correspondence, memory networks [11] to store prototypical report structures, auxilliary modules [41]–[43] for knowledge learning, or vision-language pre-training models [13] that learn broad cross-modal representations, all focusing on global alignment between entire images and complete reports. Recent works have begun to recognize the hierarchical nature of medical reports, with Jing *et al.* [6] generating reports section by section without establishing cross-modal hierarchical correspondence. Some approaches leverage multi-granularity information through auxiliary tasks, such as Liu *et al.* [23], who employ sentence-level contrastive learning, though these methods require additional annotations or focus on specific granularity levels rather than comprehensive hierarchical alignment. Wang *et al.* [44] focus on generalized medical representation learning across multiple domains using cross-modal alignment at different granularities, but their approach targets general representation learning rather than report generation specificity. Despite these advances, existing Image-Report-Only methods either process granularities sequentially, focus on single-modal hierarchies, or require auxiliary supervision, with none achieving simultaneous cross-modal hierarchical alignment across all semantic levels, creating a fundamental gap between current alignment paradigms and the hierarchical nature of clinical reasoning that our work addresses.

C. Cross-modal alignment

A core challenge in multimodal learning lies in managing the inherent heterogeneity among diverse data modalities [45]. This heterogeneity is characterized by structural differences in data organization, distributional variations in feature spaces, and semantic disparities in information representation. Existing approaches fall into several major paradigms. The most prevalent approach to addressing modality heterogeneity is shared representation learning, which aims to map multiple modalities into a common latent space. CLIP-based methods [46] exemplify this strategy through contrastive learning on large-scale image-text pairs, while recent advancements such as Uni-Code [47] enhance alignment precision via cross-modal disentanglement. However, unifying modalities in a shared space often suppresses modality-specific characteristics. Transformer-based approaches address this through cross-attention mechanisms [48], with techniques like hierarchical attention networks [49] improving semantic granularity. Yet these methods typically demand substantial computational resources and struggle with highly disparate modality pairs. Modality translation represents an alternative direction, where explicit mappings between modalities are learned through cross-modal generation [50], cycle-consistency constraints [51], or hybrid encoder-decoder architectures [52]. While effective for modality conversion, these methods risk introducing artifacts when handling complex, high-dimensional data. Cross-modal knowledge distillation has also emerged to address modality imbalance through dynamic knowledge transfer [53] and unified self-distillation frameworks [54], enhancing cross-modal coherence while often relying heavily on teacher model quality. Across these paradigms, a consistent dilemma persists: methods either prioritize cross-modal alignment at the expense of modality specificity, or preserve modality characteristics while sacrificing integration effectiveness. To address this dilemma, we propose a unified framework that achieves multi-dimensional alignment from the perspective of distribution. By conducting alignment at the distribution level, our approach simultaneously preserves modality-specific information while achieving robust cross-modal integration.

III. Method

In this section, we present the RIHA model. Section III-A provides an overview of radiology report generation, covering its mathematical formulation, basic network structure, and distance calculation method. The detailed model architecture, including each designed module, is discussed in Section III-B. Finally, the learning objectives are defined in Section III-C.

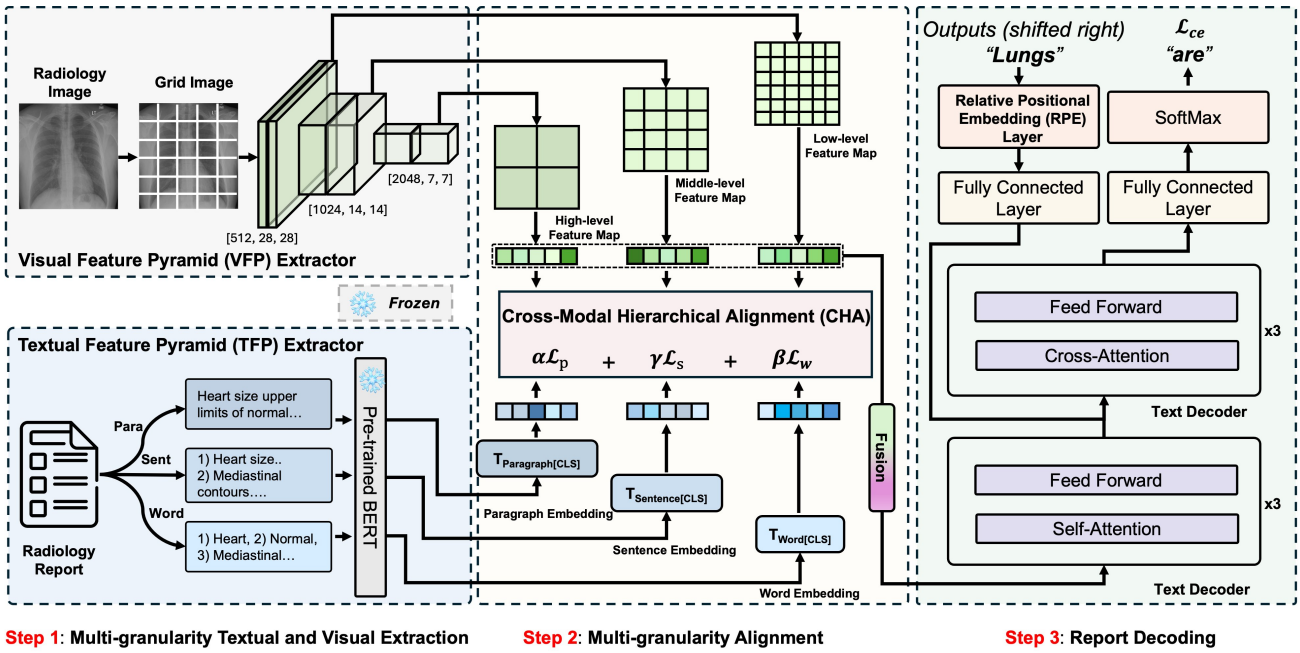

Fig. 2. The architecture of RIHA: An image is fed into the VFP Extractor to obtain shallow, middle, and high-level features. The multi-granularity text features of paragraph, sentence, and word-level features are extracted by the TFP extractor. Multi-granularity visual and textual features are then sent into CHA for hierarchical alignment. After that, refined visual and textual features are fed into a transformer encoder-decoder structure for report generation.

A. Overview and Preliminaries

Radiology Report Generation. Let us define I as a radiology image, and the goal of RRG is to generate a report R from I. Existing SOTA methods mainly utilize the encoder-decoder structure to generate reports. In practice, the visual encoder (*e.g.,* ResNet101) takes the X-ray images to extract visual features $F_I \in \mathbb{R}^{H\times W\times C}$, and then these features are flattened into a sequence of visual tokens $\hat{F}_I \in \mathbb{R}^{HW\times C}$, where the H, W, C denotes the hight, width and the number of channels, respectively. The process can be formulated as follows,

$$\mathcal{F}_{ve}(I) = \{\hat{f}_1^i, \hat{f}_2^i, ..., \hat{f}_p^i, ..., \hat{f}_N^i\} \tag{1}$$

where $\hat{f}_p^i$ represents the p-th patch visual feature in $\hat{F}_I$, and $N = H \times W$. The visual extractor is denoted as $\mathcal{F}_{ve}$

After obtaining these visual features, they are sent to the transformer encoder-decoder framework to generate a report R. The generation is an auto-regressive procedure. In particular, we can obtain the word embeddings $e^w \in \mathbb{R}^{1\times D}$ via an embedding layer for each word w in Report. The visual tokens are transferred into intermediate representations $e^i \in \mathbb{R}^{1\times D}$, where D indicates the dimensions of hidden states. Following that, the word embeddings before time stamp T combined with visual embeddings are input into the decoder for the generation of the word at the time stamp T. The whole process can be expressed in equations 3, 4 and 5.

Wasserstein Distance. The Wasserstein distance, which is also referred to as Earth Mover's Distance (EMD) [55], [56], measures the distance between two probability distributions or collections of weighted objects. This metric is derived from the solution to the classical transportation problem in operations research [57]. In mathematical terms, consider a system consisting of n suppliers $S = \{s_1, s_2, ..., s_n\}$, where each element s_i represents the supply capacity of the i-th source. These suppliers must distribute resources to m customers with fixed demand quantities $Q = \{q_1, q_2, ..., q_m\}$, where q_j denotes the requirement of the j-th consumer. The transportation cost per unit from supplier i to customer j is defined as c_{ij}. The optimization objective involves determining an optimal transportation plan $P = \{p_{ij} | I \in [1, n], j \in [1, m]\}$ that minimizes total cost while satisfying all consumer demands. The formulations can be found in equation 2.

$$\begin{aligned} \underset{p_{ij}}{\text{minimize}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}p_{ij} \\ \text{subject to} \quad & p_{ij} \geq 0, \quad i = 1, \ldots, n,\ j = 1, \ldots, m \\ & \sum_{j=1}^{m} p_{ij} = s_i, \quad i = 1, \ldots, n \\ & \sum_{i=1}^{n} p_{ij} = q_j, \quad j = 1, \ldots, m. \end{aligned} \tag{2}$$

To enhance the fine-grained alignment between X-ray images and their corresponding radiology reports, RIHA incorporates the Wasserstein distance as the fundamental component for cross-modal matching. Different from simple similarity measures [58]–[61], Wasserstein distance provides a flexible, distribution-aware metric that captures the optimal transport cost between heterogeneous feature spaces. Radiological visual features extracted from different regions of the X-ray and semantic embeddings derived from text, spanning from word-, sentence-, and paragraph-hierarchies, are treated as probability distributions. By measuring the Wasserstein distance between these hierarchically encoded representations, the model learns to align visual and textual content at multiple granularities, enabling robust and semantically meaningful correspondences

across modalities. This approach is particularly effective in the interpretation of long, descriptive radiology reports, where clinical observations are often distributed sporadically throughout the radiology report. To compute the optimal transportation plan P, we employ approximate algorithms [62]–[64]. In this work, we leverage the Wasserstein distance to align the reports and images.

$$\mathcal{F}_{en}(\hat{f}_1^i, \hat{f}_2^i, ..., \hat{f}_N^i) = (e_1^i, e_2^i, ..., e_N^i), \tag{3}$$

$$\mathcal{F}_{em}(w_1, w_2, ..., w_{T-1}) = (e_1^w, e_2^w, ..., e_{T-1}^w), \tag{4}$$

$$p_T = \mathcal{F}_{de}(e_1^i, e_2^i, ..., e_N^i; e_1^w, e_2^w, ..., e_{T-1}^w), \tag{5}$$

where $\mathcal{F}_{en}, \mathcal{F}_{em}$ and $\mathcal{F}_{de}$ represent the visual and embedding layer, and decoder, respectively. p_T is the predicted word at time stamp T.

B. Model Architecture

The overall framework of our proposed RIHA is illustrated in Fig. 2. It comprises three main stages: multi-granularity encoding, cross-modal alignment, and report decoding. In the first stage, X-ray images are processed through the visual feature pyramid (VFP) extractor to capture multi-level features. Simultaneously, the text feature pyramid (TFP) extractor encodes the input text at paragraph, sentence, and word levels. In the second stage, these multi-granularity textual features are then aligned with the corresponding visual features using the cross-modal hierarchical alignment (CHA) module. Notably, the CHA module is only required during training to learn hierarchical correspondences and can be removed during inference, significantly reducing computational overhead while maintaining the learned alignment benefits. Finally, for the third stage, the decoder incorporates the relative positional embedding (RPE) strategy to refine token-level alignment, enhancing overall coherence.

Step 1: Multi-granularity Visual and Textual Feature Encoding. Unlike conventional approaches that rely on limited textual and visual semantics, our method extracts multi-hierarchy visual features and disentangles long text into three levels of semantics: *i.e.,* paragraph, sentence, and word. Drawing inspiration from the pyramidal feature hierarchy, we capture multi-scale feature maps from various layers of the visual extractor to enhance the representation of visual content. The X-ray images are fed into a multi-hierarchical visual extractor VFP with a different number of patches, the higher the larger, to obtain shallow, middle, and high visual grid features denoted as $(\hat{V}_s, \hat{V}_m, \hat{V}_h)$ in module VFP shown as blue component in Fig. 2, where $\{s, m, h\} = \{5, 6, 7\}$ in ResNet and the $\hat{V}_s$ captures the feature from shallow layers while $\hat{V}_h$ represents the feature from deeper layer. For textual features, effectively encoding long text remains challenging despite its richer semantics. To address this, we employ TFP to decompose the long-text report into multiple levels of granularity: paragraph, sentence, and word. For sentence-level processing, we use NLTK's [65] sentence tokenizer to split paragraphs into individual sentences, then filter out short sentences that contain minimal semantic information and prioritize sentences containing clinical findings over background information. For word-level encoding, we employ a systematic selection process rather than processing all words indiscriminately. Specifically, we use NLTK's POS tagger to identify parts of speech and selectively extract nouns, adjectives, and quantifiers that carry the most diagnostic information in radiology reports. We further prioritize medical terminology using clinical vocabulary filters and extract key phrases commonly appearing in radiology findings (e.g., "consolidation," "effusion," "pneumothorax"). This selective approach ensures focus on clinically relevant semantic units.

Finally, a pre-trained Bio_ClinicalBERT language model is applied to encode the resulting segments, with the model frozen during training to preserve medical domain knowledge. We obtain paragraph-level semantics directly from BERT embeddings, denoted as $t_p = \{p_1, p_2, ..., p_{|t_p|}\}$, where $p_i \in \mathbb{R}^d$ represents the i-th word representation with dimension d. For sentence and word semantics, we extract representations from the <CLS> token, which aggregates contextual information from all tokens through self-attention mechanisms, effectively capturing sequence-level semantics. These are denoted as $t_s = \{s_1, s_2, ..., s_{|t_s|}\}$ and $t_w = \{w_1, w_2, ..., w_{|t_w|}\}$, where s_i and w_i indicate the i-th sentence and word representation with dimension d, respectively. For word-level encoding, <CLS> representations are extracted from sequences containing selected keywords with minimal context to maintain semantic coherence.

Step 2: Multi-granularity Alignment. To address the alignment inconsistency between text and image, multi-granularity visual and textual semantics are first extracted as described in Step 1. However, effectively aligning these cross-modal features remains a critical challenge. Traditional methods often rely on mean representation distances, which overlook important local semantic cues. To overcome this, we model the alignment as an optimal transportation problem [66], aiming to minimize the transportation cost from grid-based visual features to their corresponding textual representations. Specifically, we employ the Wasserstein distance to compute this cost, capturing the fine-grained correspondence between features. For illustration, we consider the alignment between a sentence t_s and a middle-level visual feature $\hat{V}_m$ (see Fig.3), as the alignment strategy remains consistent across different granularities. According to the definition in Sec. III-A, we take $\hat{V}_m$ as the suppliers and t_s as the customers. The transportation cost per unit can be calculated as the distance between the i-th visual grid representation and j-th sentence representation $l_{ij} = ||v_i - s_j||_2$. By minimizing this cost, we obtain the optimal transportation flow F, which can be integrated into the neural network for end-to-end training, enabling effective alignment across different semantic levels. After hierarchical alignment, the visual features from three hierarchies are fused for Step 3. Specifically, the aligned paragraph-level, sentence-level, and word-level visual features - $\hat{V}_h, \hat{V}_m, \hat{V}_s$ - are concatenated as F_{fused} and then processed through a linear projection layer to reduce dimensionality before being fed into the transformer encoder.

Step 3: Report Decoding with Relative Position. In step 2, the hierarchical alignment process extracts enhanced visual features, which are then processed by the transformer

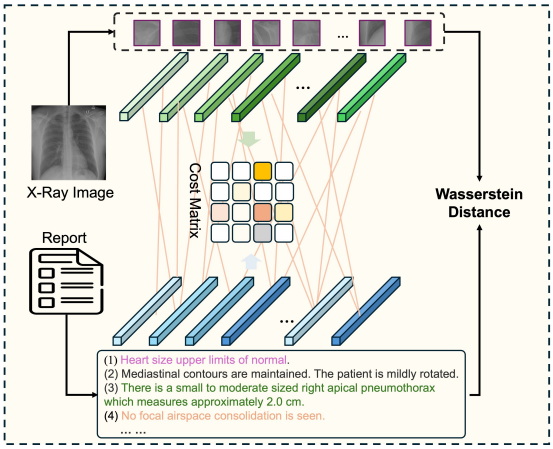

Fig. 3. An example of leveraging optimal transport to align hierarchical textual and visual features treats the grid visual features as suppliers and textual features as customers. The goal is to minimize the transportation cost, yielding an optimal alignment captured in the cost matrix.

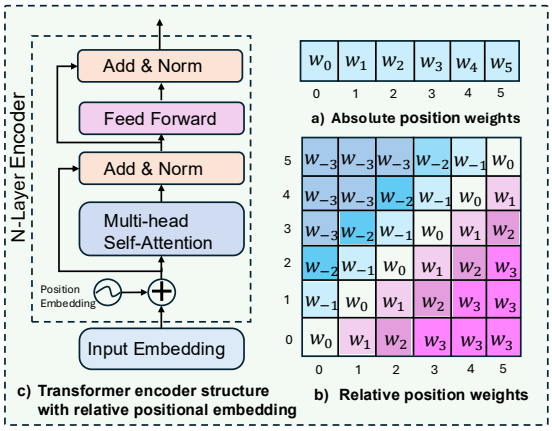

Fig. 4. An illustration comparing relative and absolute positional embeddings in transformers, where the clipped value $k = 3$ represents the maximum allowable relative position distance. a) Absolute position embedding weights. b) Relative position embedding weights. c) The transformer encoder structure with relative position embeddings. For further details, see [68].

encoder. This encoder utilizes multiple layers of multi-head self-attention and feedforward networks to capture complex visual representations. The decoder subsequently generates the radiology report, attending to both the previously generated tokens through masked self-attention and the encoded visual features via cross-attention mechanisms. To strengthen the alignment between the radiology report R and the corresponding image I at the token level, we incorporate relative position encoding (RPE) into the self-attention mechanism. Unlike the fixed positional encoding used in standard Transformers, which only captures absolute token positions, RPE effectively encodes relative positional information, directly modeling the contextual relationships between tokens as shown in Fig 4. This approach addresses the limitations of vanilla Transformers, which struggle with cross-modal alignment due to their reliance on fixed, sequence-based indices. The relative position representation and more details can be referred to [67].

C. Learning Objectives

The overall objective function of our model includes CrossEntropy loss $\mathcal{L}_{ce}$ and hierarchical alignment loss $\mathcal{L}_{alignment}$. According to the calculation in Sec. III-B, the hierarchical alignment loss could be further separated into paragraph loss $\mathcal{L}_p$ (*i.e.,* V_h and paragraph t_p), sentence loss $\mathcal{L}_s$ (*i.e.,* V_m and sentence t_s), and word loss $\mathcal{L}_w$ (*i.e.,* V_w and word t_w). We introduce the weights α, γ, β to balance the contributions of these components. Therefore, the overall learning objectives for the proposed model are as follows.

$$\mathcal{L}_{alignment} = \alpha\mathcal{L}_p + \gamma\mathcal{L}_s + \beta\mathcal{L}_w, \tag{6}$$

$$\mathcal{L} = \mathcal{L}_{ce} + \mathcal{L}_{alignment} \tag{7}$$

IV. Experiments

A. Datasets and Evaluation Metrics

IU-Xray [69] is a widely used chest X-ray dataset containing 3,996 reports from the Indiana Network and 8,121 associated frontal or lateral view images from hospital picture archiving systems. Each report is composed of three sections: indication, findings, and impression. In practical experimental settings [11], [70], [71], only the findings section is used as the target text for generation based on the paired images. Reports lacking a findings section are excluded, reducing the actual dataset to 2,955 reports. For training, the dataset is typically split into 2,069 training, 296 validation, and 590 testing samples, with no overlap between subsets, as each report corresponds to a unique patient. Additionally, previous studies have standardized the maximum report length to 60 tokens.

MIMIC-CXR [72] is a comprehensive dataset comprising 227,835 imaging studies from 65,379 patients, collected at the Beth Israel Deaconess Medical Center Emergency Department between 2011 and 2016. Each report includes sections such as Indication, Preamble, Comparison, Findings, and Impression. In this work, we focus on generating the Findings section as the RRG task. The dataset provides 473,057 X-ray images and 206,563 corresponding reports. For consistency with prior studies, we adopt the official data split: 152,173 reports and 270,790 images for training, 1,196 reports and 2,130 images for validation, and 2,347 reports and 3,858 images for testing. Unlike the IU-Xray dataset, we use single X-ray images paired with their respective reports, as approximately 35% of the MIMIC-CXR reports are associated with only one image. The maximum report length is set to 100 tokens.

Evaluation Metrics for RRG encompass two main aspects: natural language generation (NLG) and clinical efficacy (CE). NLG performance is measured using BLEU [73], METEOR [74], ROUGE-L [75], and CIDEr-D [76]. Clinical efficacy is assessed via the CheXpert tool [9], which labels each report based on the presence of 14 findings, where 1, 0, -1, and *blank* denote confidently present, confidently absent, uncertainly present, and not mentioned, respectively. Consistent with prior studies, we classify 1 and -1 as positive labels, while 0 and *blank* are considered negative. We then compute the mean precision, recall, and F1 score across the 14

classes using these labels from both the predicted and ground truth reports.

B. Implementation Details

The model is trained end-to-end using a PyTorch implementation on a single NVIDIA A6000 GPU. Before feeding images into the model, we crop the images into 224×224 pixels, and use the ResNet101 pretrained on ImageNet as the backbone of the visual extractor. The final output from the visual extractor is a 7×7 feature map with the dimension of each feature set to 2048. For extracting the hierarchical textual feature, we first leverage the natural language toolkit (*i.e.,* NLTK) to split the paragraph into sentence- and word-level information and save it to independent configuration files. Then, we fed three levels of textual content into a pre-trained BERT model, where we employ the Bio_ClinicalBERT, to extract the hierarchical textual features. For the encoder-decoder backbone, we employ a randomly initialized memory-driven Transformer [11] with 3 layers and 8 attention heads, and 512 dimensions for the hidden states. The hyperparameter settings in Wasserstein distance, including the regularization coefficient σ and the marginal penalization τ, are set as 0.1 and 0.5, respectively. The value for each hyperparameter depends on the performance of the model on the validation subset. During the training stage, the learning rate is set as $1e-3$ for the visual extractor and $2e-3$ for the encoder-decoder on the IU-Xray dataset, while the values are $5e-5$ and $1e-4$ on the MIMIC-CXR dataset. We use Adam to optimize our model. In addition, the training epoch and batch size are set as 30 and 64, respectively.

C. Comparison to SOTA Methods

In this section, we compare our method with a wide range of methods, including the conventional image captioning [28], [33] as well as previous RRG approaches [10]–[12], [14], [70], [71], [77]–[79]. Most of the methods evaluate their performance on the two public datasets - IU-Xray and MIMIC-CXR - and share similar experimental settings as those used in work [11], so we cite some experimental results directly from their papers. In the comparisons, ATT2IN [33] applies a reinforcement learning strategy originally designed for image captioning to report generation. R2Gen [11] introduces a relational memory mechanism to capture key information for automatic radiology report generation. R2GenCMN [14] further advances this approach with a cross-modal memory network, enhancing multimodal interactions. XProNet [10] constructs a cross-modal prototype network to strengthen cross-modal pattern learning, thereby improving report accuracy. RGRG [78] adopts an interactive, explainable approach by detecting anatomical regions and generating localized descriptions. M2KT [20] and GSKET [21] enhance radiology report generation by incorporating medical knowledge: the former learns a medical knowledge base while performing multi-modal alignment, and the latter incorporates both general medical knowledge and disease-specific knowledge. CAMANET [12] employs class activation maps to guide attention mechanisms for generating more accurate and visually-grounded radiology reports from medical images. Finally, PromptMRG [79] leverages diagnosis-driven prompts, transforming disease classification results into token prompts that guide the report generation process.

As shown in Tab I, our proposed RIHA method outperforms most of the state-of-art approaches on two benchmark datasets. For the IU-Xray dataset, RIHA achieves superior performance across multiple metrics, obtaining the highest scores in BLEU-1 (0.498), BLEU-4 (0.218), METEOR (0.204), and CIDEr (0.421). Notably, our method outperforms recent strong baselines including M2KT [20] by 1.8% on BLEU-1 and 25.3% on BLEU-4, GSKET [21] by 1.6% on BLEU-1 and 23.2% on BLEU-4, CAMANET [12] by 4.2% on BLEU-1 and 21.8% on BLEU-4 and PromptMRG [79] by 2.0% on BLEU-1 and 15.3% on BLEU-4. On the more challenging MIMIC-CXR dataset, RIHA demonstrates competitive results, achieving the best scores in BLEU-1 (0.385), BLEU-2 (0.251), METEOR (0.159), and ROUGE-L (0.293). Our method shows consistent improvements over recent approaches including M2KT [20] (4.6% improvement on BLEU-1), CAMANET [12] (4.9% improvement on BLEU-1), and COMG [80] (11.6% improvement on BLEU-1). These results demonstrate the effectiveness of our proposed method in generating accurate and semantically meaningful radiology reports across different datasets.

D. Clinical Efficacy

To assess the effectiveness of our proposed RIHA method in capturing abnormal clinical information, we compare its performance with several prior methods on the MIMIC-CXR dataset. The Clinical Efficacy (CE) metrics used in this evaluation were first introduced in R2Gen [11]. For MIMIC-CXR, medical term extraction relies on the rule-based CheXpert labeler, which automatically generates labels for both reference and generated reports. Since the IU-Xray dataset lacks standardized labels, CE metrics are reported exclusively for MIMIC-CXR. As presented in Tab. II, our RIHA approach achieves the highest F1-Score of 0.375, surpassing R2GenCMN [14] (0.374) and significantly outperforming XProNet (0.353) and PromptMRG [79] (0.352). In terms of precision, RIHA achieves 0.486, outperforming all baselines, including PromptMRG (0.471) and XProNet [10] (0.463). For recall, RIHA attains 0.298, outperforming XProNet (0.285) and PromptMRG (0.198), though it falls short of R2GenCMN's recall (0.325). These results highlight RIHA's ability to balance precision and recall, demonstrating its potential for accurate medical condition identification in clinical practice.

E. Ablation Studies

In this section, we evaluate the effectiveness of our proposed modules, including the visual feature pyramid (VFP), cross-modal hierarchical alignment (CHA), and relative positional embedding (RPE). We also provide a detailed analysis of hyperparameter settings for hierarchical alignment loss, the choice of textual pre-trained models, and distance calculation methods used in the alignment loss.

TABLE I
COMPARISON WITH STATE-OF-THE-ART METHODS ON IU-XRAY AND MIMIC-CXR DATASETS. OPTIMAL AND SUBOPTIMAL PERFORMANCE IS HIGHLIGHTED.

DATASET	METHOD	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDER
IU-Xray [69]	ADDAATT [28]	0.220	0.127	0.089	0.068	-	0.308	0.295
	ATT2IN [33]	0.224	0.129	0.089	0.068	-	0.308	0.220
	HRGR [70]	0.438	0.298	0.208	0.151	-	0.322	0.343
	COAT [6]	0.455	0.288	0.205	0.154	-	0.369	0.277
	R2GEN [11]	0.390	0.293	0.236	0.161	0.187	0.371	-
	CMCL [77]	0.473	0.305	0.217	0.162	0.186	0.378	-
	PPKED [71]	0.483	**0.315**	0.224	0.168	0.190	0.376	0.351
	R2GENCMN [14]	0.402	0.310	0.225	0.175	0.190	0.365	0.344
	XPRONet [10]	0.480	0.312	0.225	0.175	0.190	0.364	-
	RGRG [78]	0.373	0.249	0.175	0.126	0.168	0.264	-
	M2KT [20]	0.489	**0.315**	0.228	0.174	-	**0.379**	**0.405**
	GSKET [21]	**0.490**	0.311	0.236	0.177	-	0.367	0.379
	COMG [80]	0.481	0.313	0.232	0.183	0.189	0.371	-
	EKAGen [81]	0.471	0.308	0.223	0.172	0.201	0.368	-
	CAMANET [12]	0.478	0.314	0.229	0.179	**0.202**	0.361	0.412
	PromptMRG [79]	0.488	0.253	**0.237**	**0.189**	0.192	0.374	-
	RIHA (**Ours**)	**0.498**	**0.317**	**0.239**	**0.218**	**0.204**	**0.376**	**0.421**
MIMIC-CXR [72]	ST [82]	0.299	0.184	0.121	0.084	0.124	0.263	-
	ATT2IN [33]	0.325	0.203	0.136	0.096	0.134	0.276	-
	ADAATT [28]	0.299	0.185	0.124	0.088	0.118	0.266	-
	TOPDOWN [29]	0.317	0.195	0.130	0.092	0.128	0.267	-
	R2GEN [11]	0.353	0.218	0.145	0.103	0.142	0.270	-
	CMCL [77]	0.344	0.217	0.140	0.097	0.133	0.281	-
	R2GENCMN [14]	0.348	0.206	0.135	0.094	0.136	0.266	0.158
	PPKED [71]	0.360	0.224	0.149	0.106	**0.149**	**0.284**	**0.237**
	XPRONet [10]	0.344	0.215	0.146	0.105	0.138	0.279	0.154
	M2KT [20]	0.368	0.220	0.149	**0.107**	-	0.259	0.110
	COMG [80]	0.345	0.214	0.143	0.104	0.136	0.275	-
	EKAGen [81]	0.365	**0.233**	**0.152**	0.105	0.143	0.271	-
	CAMANET [12]	0.367	0.219	0.151	**0.107**	0.142	0.279	0.158
	PromptMRG [79]	**0.369**	0.223	0.142	0.099	0.147	0.260	-
	RIHA (**Ours**)	**0.385**	**0.251**	**0.162**	**0.109**	**0.159**	**0.293**	**0.161**

TABLE II
CLINICAL EFFICACY COMPARISONS ON MIMIC-CXR DATASET. OPTIMAL PERFORMANCE IS HIGHLIGHTED.

Method	Precision	Recall	F1-Score
ATT2IN [33]	0.268	0.203	0.204
ADDAATT [28]	0.322	0.239	0.249
R2Gen [11]	0.406	0.213	0.280
R2GenCMN [14]	0.440	0.325	0.374
XProNet [10]	0.463	0.285	0.353
PromptMRG [79]	0.471	0.198	0.352
RIHA(**Ours**)	**0.486**	**0.298**	**0.375**

1) Effectiveness of proposed modules: The following models are evaluated in our experiment.

- **Base**: The base model only includes the basic visual extractor and the basic transformer encoder-decoder structure.
- **Base+VFP**: The visual features are extracted from the 5-th, 6-th and 7-th layer of the basic visual extractor, and then are fused to the subsequent module.
- **Base+VFP+TFP+CHA**: This introduces the multi-granularity alignment by the CHA module between visual features from different layers and textual features from paragraph, sentence, and word levels.
- **Base+VFP+TFP+CHA+RPE**: This is the full RIHA containing all our proposed components.

As shown in Tab. III, we progressively integrated three critical components into our base model: Visual Feature Projection (VFP), Cross-modal Hierarchical Attention (CHA), and Relative Positional Embedding (RPE). Each addition demonstrated notable performance gains across multiple metrics. On the IU-Xray dataset, incorporating VFP alone increased the BLEU-1 score from 0.390 to 0.419. Adding CHA further boosted scores (BLEU-1: 0.492, BLEU-4: 0.178), while the full model, including RPE, achieved the highest performance (BLEU-1: 0.498, CIDEr: 0.421, METEOR: 0.207). Similar improvements were observed on the more challenging MIMIC-CXR dataset, where each component incrementally enhanced the overall results, culminating in the best scores for the complete model (BLEU-1: 0.385, ROUGE-L: 0.293). These results underscore the distinct and complementary contributions of each component to overall report generation quality.

2) Analysis on hyperparameters for alignment loss: We conduct extensive experiments to analyze the sensitivity of the hyper-parameters α, γ, and β at the paragraph, sentence, and word levels, respectively, across various evaluation metrics (see Tab. IV). The results show that the configuration (0.5, 0.2, 0.3) achieves the best performance, with BLEU-1, BLEU-4, METEOR, and ROUGE-L scores of 0.498, 0.176, 0.207, and 0.368, respectively. We observe that increasing the paragraph-level weight (α) from 0.3 to 0.5 consistently im-

TABLE III
THE EXPERIMENTAL RESULTS OF ABLATION STUDIES ON THE IU-XRAY AND MIMIC-CXR DATASETS. OPTIMAL PERFORMANCE IS HIGHLIGHTED.

IU-Xray	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDER**
Base	0.390	0.237	0.161	0.113	0.170	0.322	0.336
+VFP	0.419	0.262	0.184	0.137	0.172	0.340	0.415
+VFP+TFP+CHA	0.492	0.313	0.231	0.178	0.204	0.369	0.418
+VFP+TFP+CHA+RPE	**0.498**	**0.317**	**0.239**	**0.218**	**0.204**	**0.376**	**0.421**
MIMIC-CXR	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDER**
Base	0.363	0.225	0.152	0.101	0.142	0.280	0.144
+VFP	0.371	0.226	0.155	0.106	0.152	0.286	0.150
+VFP+TFP+CHA	0.382	0.248	0.158	0.108	0.158	0.292	0.159
+VFP+TFP+CHA+RPE	**0.385**	**0.251**	**0.162**	**0.109**	**0.159**	**0.293**	**0.161**

TABLE IV
MODEL PERFORMANCE UNDER DIFFERENT HYPERPARAMETERS (α, γ, β) ON THE IU-XRAY DATASET. OPTIMAL PERFORMANCE IS HIGHLIGHTED.

(α, γ, β)	**BLEU-1**	**BLEU-4**	**METEOR**	**ROUGE-L**
(0.3, 0.4, 0.3)	0.437	0.162	0.187	0.340
(0.4, 0.2, 0.4)	0.483	0.172	0.201	0.362
(0.5, 0.2, 0.3)	**0.498**	**0.218**	**0.204**	**0.376**
(0.6, 0.2, 0.2)	0.471	0.169	0.198	0.361
(0.7, 0.2, 0.1)	0.468	0.165	0.191	0.359
(1.0, 0.0, 0.0)	0.449	0.160	0.181	0.354
(0.0, 1.0, 0.0)	0.429	0.151	0.173	0.347
(0.0, 0.0, 1.0)	0.425	0.136	0.170	0.343

TABLE V
EFFECT OF USING DIFFERENT BERT PRE-TRAINED MODELS TO EXTRACT THE TEXTUAL FEATURES ON THE IU-XRAY DATASET. OPTIMAL PERFORMANCE IS HIGHLIGHTED.

BERT-type	**BLEU-1**	**BLEU-4**	**METEOR**	**ROUGE-L**
bert-base-uncased	0.487	0.198	0.196	0.349
PubMedBERT-base	0.496	0.213	0.204	0.371
Bio_ClinicalBERT	**0.498**	**0.218**	**0.204**	**0.376**

proves performance, highlighting the importance of paragraph-level coherence. However, further increasing α beyond 0.5 leads to diminishing returns, indicating a trade-off between paragraph-level emphasis and sentence/word-level features. Extreme cases where one component is maximized (e.g., $\alpha = 1.0$, $\gamma = 1.0$, or $\beta = 1.0$) result in significantly lower performance, confirming that multi-level integration is crucial for accurate text evaluation. This supports our hypothesis that focusing exclusively on any single linguistic level overlooks the complex interdependencies that drive overall text quality.

3) Analysis on textual pre-trained models: To extract multi-granularity textual features, we use the pre-trained BERT model, which remains frozen during training. While the base BERT model is trained on natural language, Bio_ClinicalBERT and PubMedBERT are fine-tuned on clinical text to capture domain-specific medical language patterns. Intuitively, domain-specific BERT models should yield better results. Tab. V compares the performance of different BERT architectures on the IU-Xray dataset. The results show a clear advantage for domain-specific models. The general-purpose bert-base-uncased model performs reasonably well (BLEU-1: 0.487, BLEU-4: 0.198), but domain-specific models show significant improvements across all metrics. PubMedBERT-base, pre-trained on biomedical literature, improves BLEU-4 (0.213) and ROUGE-L (0.371) scores. Notably, Bio_ClinicalBERT, pre-trained on clinical text, outperforms all models (BLEU-1: 0.498, BLEU-4: 0.218, METEOR: 0.204, ROUGE-L: 0.376). These findings confirm that domain-specific pretraining effectively captures medical terminology and semantic relationships, improving the accuracy and relevance of radiology report generation.

TABLE VI
EFFECT OF USING DIFFERENT DISTANCE CALCULATION METHODS FOR THE PROPOSED CHA MODULE ON IU-XRAY DATASET. OPTIMAL PERFORMANCE IS HIGHLIGHTED.

Distance Calculation	**BLEU-1**	**BLEU-4**	**METEOR**	**ROUGE-L**
Wasserstein distance	**0.498**	**0.218**	**0.204**	**0.376**
L_1 distance	0.487	0.209	0.199	0.368
L_2 distance	0.482	0.205	0.197	0.364
Cosine distance	0.491	0.213	0.201	0.371
KL divergence	0.476	0.201	0.195	0.360

4) Analysis on the distance calculation: To assess the impact of different distance metrics on the alignment of visual and textual data distributions, we conduct experiments on the IU-Xray dataset using various distances: Wasserstein, L_1, L_2, Cosine, and KL divergence. The results, shown in Tab. VI, indicate that Wasserstein distance consistently outperforms the others, achieving the highest scores across all metrics (BLEU-1: 0.498, BLEU-4: 0.218, METEOR: 0.204, ROUGE-L: 0.376). This superior performance is due to the Wasserstein distance's ability to capture geometric relationships between distributions, even with minimal overlap. Cosine distance also performs competitively (BLEU-1: 0.491, ROUGE-L: 0.371), highlighting that directional similarity effectively captures cross-modal relationships. Traditional L_1 and L_2 distances produce moderate results, with L_1 consistently outperforming L_2, suggesting that the Manhattan distance better suits the discrete nature of our task. KL divergence shows the lowest performance, likely due to its sensitivity to zero-probability regions and its asymmetric nature. These findings indicate that optimal cross-modal alignment is best achieved by using distance metrics that consider the structural properties of underlying distributions, rather than simply comparing proba-

bility masses.

5) Analysis on the computational cost of distance calculation methods: Due to the high computational cost of the Wasserstein distance, we conduct experiments to evaluate the computational cost between standard OT calculation and existing other optimized OT calculations in Tab. IX. Based on the table, the results demonstrate that both accelerated OT strategies substantially improve computational efficiency. Sampling-based OT achieves the greatest reduction with 43% faster training and 24% lower memory consumption, while prototype-based OT provides a 32% speedup in training time and 20% reduction in GPU memory usage compared to standard OT. In this work, we employed standard OT to thoroughly investigate its suitability and effectiveness for hierarchical alignment in radiology report generation. Future work will explore how optimized OT solvers can further enhance both computational efficiency and model performance.

6) Analysis on the backbone of visual extractor: To validate our choice of ResNet101 as the visual backbone, we conduct comparative experiments with alternative architectures on the IU-Xray dataset. As shown in table VIII, ResNet101 achieves the best performance across all metrics, with BLEU-1 of 0.498, BLEU-4 of 0.218, METEOR of 0.204, and ROUGE-L of 0.376. DenseNet121 demonstrates competitive but slightly lower performance (BLEU-1: 0.485, BLEU-4: 0.206), while ViT shows the weakest results (BLEU-1: 0.479, BLEU-4: 0.201). The superior performance of ResNet101 can be attributed to its robust feature extraction capability for medical images and well-established pre-training on large-scale datasets. Given the relatively small size of medical datasets, CNN-based architectures like ResNet101 demonstrate better generalization than transformer-based models, which typically require larger training corpora. These results justify our architectural choice and align with previous findings in the radiology report generation literature.

7) Analysis on the frozen and finetuned BERT encoder: To investigate the impact of BERT encoder training strategy on report generation performance, we compared frozen and fine-tuned Bio_ClinicalBERT encoders on the IU-Xray dataset. As shown in the table, the frozen BERT encoder achieves comparable or slightly better performance across all metrics (BLEU-1: 0.498, BLEU-4: 0.218, METEOR: 0.204, ROUGE-L: 0.376) compared to the fine-tuned version (BLEU-1: 0.495, BLEU-4: 0.215, METEOR: 0.202, ROUGE-L: 0.373). The marginal performance difference (0.5-1.5% across metrics) suggests that fine-tuning the pre-trained BERT encoder provides minimal additional benefit for our hierarchical alignment framework. This result can be attributed to several factors. First, Bio_ClinicalBERT has already been extensively pre-trained on clinical text, capturing domain-specific medical terminology and semantic relationships effectively. Second, freezing the BERT encoder prevents potential overfitting and forgetting to the relatively small training corpus and preserves the rich linguistic knowledge learned during pre-training. Based on these findings, we adopt the frozen BERT encoder strategy in our final model, which offers the advantages of reduced computational cost, faster training convergence, while maintaining competitive performance.

TABLE VII
EFFECT OF USING DIFFERENT FROZEN OR FINE-TUNED BERT TO EXTRACT THE TEXTUAL FEATURES ON THE IU-XRAY DATASET.

BERT encoder	BLEU-1	BLEU-4	METEOR	ROUGE-L
Frozen	0.498	0.218	0.204	0.376
Fine-tuned	0.495	0.215	0.202	0.373

TABLE VIII
EFFECT OF USING DIFFERENT VISUAL ENCODERS TO EXTRACT THE VISUAL FEATURES ON THE IU-XRAY DATASET. OPTIMAL PERFORMANCE IS HIGHLIGHTED.

Visual Encoder	BLEU-1	BLEU-4	METEOR	ROUGE-L
ResNet101	**0.498**	**0.218**	**0.204**	**0.376**
DenseNet121	0.485	0.206	0.197	0.368
ViT	0.479	0.201	0.193	0.362

TABLE IX
COMPUTATIONAL COST OF DIFFERENT DISTANCE CALCULATION METHODS ON THE IU-XRAY DATASET.

Distance Strategy	Training Time(h)	GPU Memory Usage(GB)
L_1 Distance	1.2	8.4
L_2 Distance	1.3	8.6
Cosine Similarity	1.4	8.8
Standard OT	2.8	12.1
Sampling-based OT	1.6	9.2
Prototype-based OT	1.9	9.7

F. Qulitative Results

To further assess the effectiveness of our proposed method, we present qualitative results in Fig. 5, Fig. 6, and Fig. 7. Fig. 5 highlights the superior performance of our RIHA method in generating radiology reports. Unlike the baseline model, which incorrectly identifies “confluent consolidation” in the anteroposterior (AP) view, RIHA accurately detects “clear of large confluent consolidation” and appropriately evaluates the cardiac silhouette, aligning closely with ground-truth annotations. Similarly, in posteroanterior (PA) and lateral views, RIHA correctly identifies the absence of pleural effusion and reports “mild infectious process, specifically no evidence of focal consolidation pneumothorax or pleural effusion,” reflecting a more precise assessment. Fig. 6 further demonstrates the incremental improvements from our architectural components. The base model with Visual Feature Pyramid (VFP) extractor shows enhanced detection of bilateral lower lobe collapse, while adding Cross-Hierarchy Attention (CHA) enables more accurate identification of pulmonary vascular congestion and atelectasis. The complete model (Base+VFP+CHA+RPE) achieves comprehensive analysis, accurately detecting pleural effusions and providing clearer evaluations of cardiac and mediastinal contours. Finally, Fig. 7 presents attention maps,

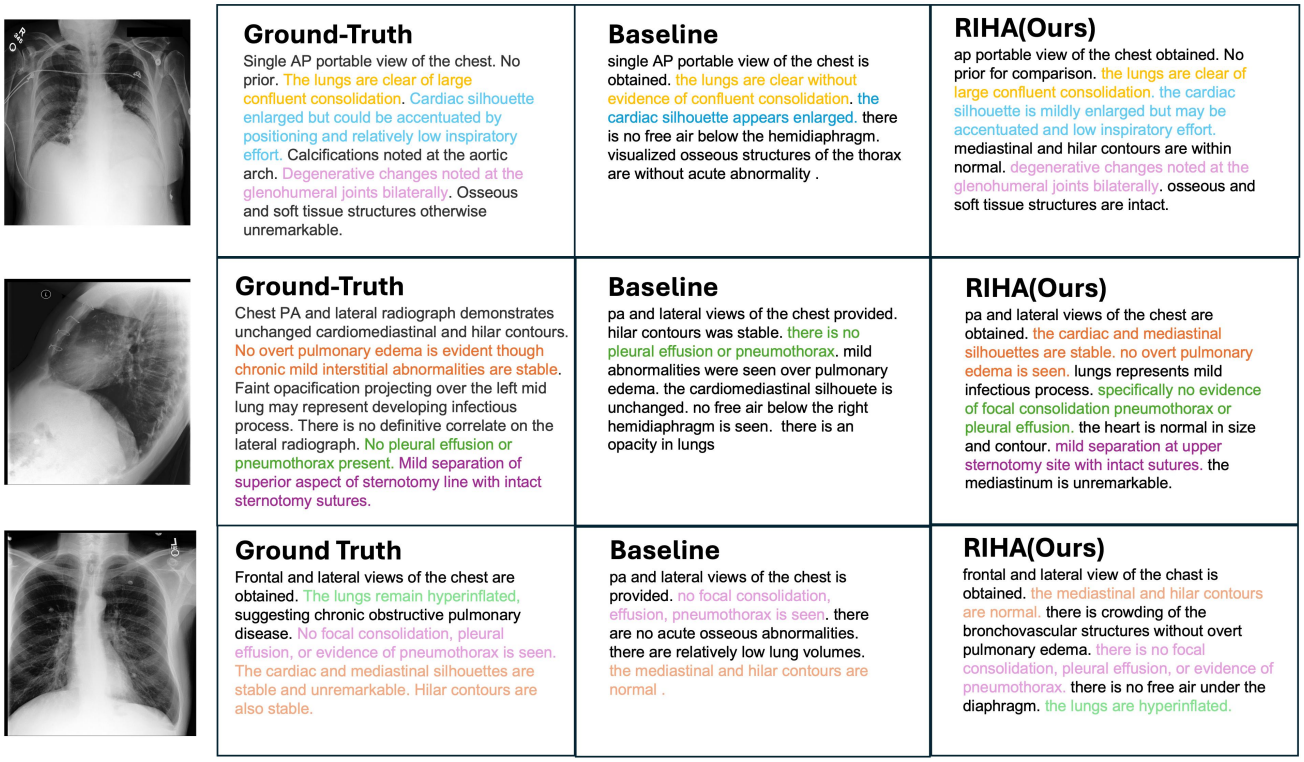

Fig. 5. Examples of generated reports from the MIMIC-CXR testing subset using the baseline model and our proposed RIHA method. Identical findings in the ground truth (GT) and generated reports are highlighted with matching colors, demonstrating the superior performance of our approach.

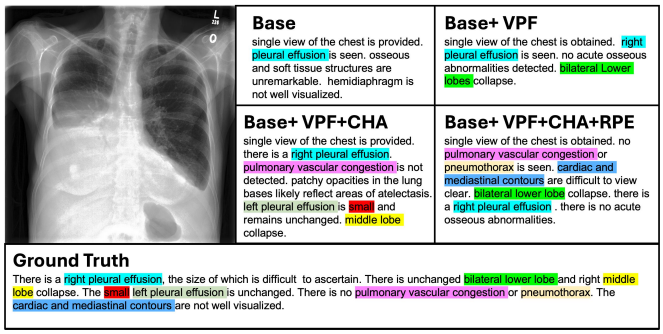

Fig. 6. Example reports generated by incrementally incorporating the proposed modules into the baseline model are shown. Key medical terms are highlighted in different colors to clearly differentiate model performance.

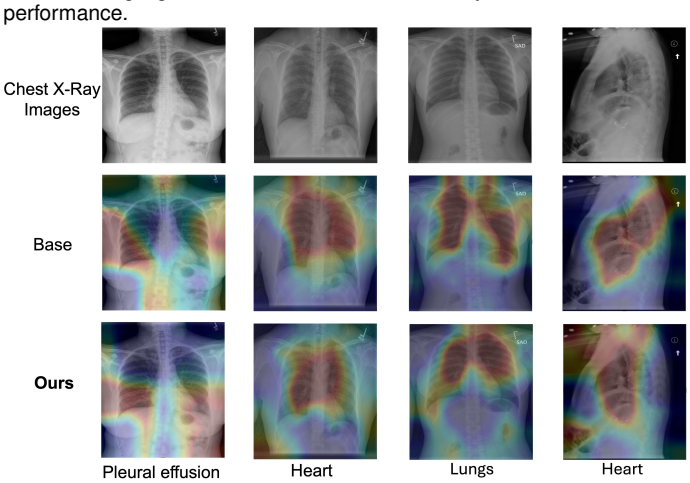

Fig. 7. Attention map visualizations for various keywords from the Baseline and RIHA models reveal that RIHA assigns more precise attention regions, highlighting its improved focus for each keyword.

illustrating our model's capacity to align textual descriptions with corresponding anatomical regions. These heat maps show precise localization of pleural effusion, cardiac structures, and lung fields, validating RIHA's ability to effectively learn the spatial relationships between radiological findings and their textual descriptions. This visual evidence supports our quantitative results, confirming RIHA's potential for generating expert-level radiological interpretations.

G. Clinical Implications

1) Clinical analysis by LLM and radiologists: To assess the clinical interpretability and quality of generated reports, we conduct a human expert evaluation study. We randomly sample 30 cases from the test set and generate reports using Baseline, R2Gen, XProNet, and our proposed RIHA model. A board-certified radiological expert from Tan Tock Seng Hospital (TTSH) independently evaluated all generated reports across multiple dimensions, including clinical correctness, coherence, completeness, and diagnostic accuracy, providing an overall score for each model. As shown in the Tab. XI, RIHA achieves the highest clinician score of 8.0/10, substantially outperforming XProNet (6.7), R2Gen (6.3), and the Baseline model (5.8). Notably, RIHA's score demonstrates the closest alignment between automated LLM evaluation (8.2) and clinical expert judgment, with a minimal gap of only 0.2 points, compared to larger discrepancies observed in other models (0.4-0.5 points). This strong concordance indicates that RIHA-generated reports better capture the clinical reasoning patterns and diagnostic standards expected by a radiological expert.

2) Clinical error analysis: To evaluate the clinical reliability of our approach beyond standard NLG metrics, we conduct a comprehensive analysis of critical finding detection accuracy. Using the CheXpert [9] labeler to automatically annotate both GT and generated reports, we perform a focused case study on cardiomegaly detection in the IU-Xray test set - a critical cardiac finding that significantly impacts patient management decisions. We categorize errors into three types: Omission (O), where a finding present in GT is absent from the generated report; Misclassification (M), where GT and generated reports have conflicting conclusions; and Hallucination (H), where a finding appears in the generated report but not in GT. Based on the definitions, M encompasses both O and H, i.e., M = O + H. As shown in Tab. X, RIHA achieves superior clinical accuracy with a 94.8% detection rate and only 5.2% misclassification rate for cardiomegaly cases, comprising 3.1% omission and 2.1% hallucination errors. This represents a substantial improvement over baseline methods (79% detection, 21% misclassification) and SOTA approaches including R2GenCMN (81% detection, 19% misclassification) and XProNet (86.2% detection, 13.8% misclassification). Notably, RIHA demonstrates the lowest omission rate (3.1%) among all methods, indicating its effectiveness in capturing critical findings. The significantly reduced error rates across all categories demonstrate the clinical value of our hierarchical alignment approach in detecting subtle but clinically significant abnormalities.

TABLE X
CARDIOMEGALY DETECTION AND ERROR RATES BY DIFFERENT MODELS. THE CHARACTERS O, M, AND H REPRESENT OMISSION, MISCLASSIFICATION, AND HALLUCINATION, RESPECTIVELY.

Model	Cardiomegaly detected	M	O	H
RIHA	94.8%	5.2%	3.1%	2.1%
Baseline	79.0%	21.0%	14.5%	6.5%
R2GenCMN	81.0%	19.0%	13.1%	5.9%
XProNet	86.2%	13.8%	9.2%	4.6%

TABLE XI
EVALUATION SCORE OF THE GENERATED REPORTS FROM DIFFERENT MODELS BY LLM AND CLINICIANS.

Model	Baseline	R2Gen	XProNet	RIHA
LLM Score	6.2	6.8	7.1	8.2
Clinician Score	5.8	6.3	6.7	8.0

V. LIMITATIONS AND DISCUSSIONS

Despite its superior performance in radiology report generation, our current model has some limitations. First, it relies heavily on the manual disentanglement of three granularities, a step that could be streamlined in the future by leveraging large language models for automatic multi-granularity selection and disentanglement. We could train lightweight segmentation modules to identify paragraph and sentence boundaries by leveraging the consistent structure of radiology reports. Alternatively, we could utilize medical-specific large language models to automatically extract semantically meaningful keywords while maintaining clinical relevance. Second, the optimal transport algorithm used for alignment is computationally intensive, but this can be mitigated by enhancing the original algorithm, for instance, through optimized sampling or prototype-based refinement. Nonetheless, our model remains efficient and lightweight, with the hierarchical alignment module being removable during testing, significantly boosting inference speed. This modular design also makes the approach adaptable to related tasks, such as paragraph generation, text-to-image, or image-to-text generation. Third, our current framework handles uncertainty expressions (e.g., "mild," "possible," "likely") implicitly through the pre-trained Bio_ClinicalBERT encoder rather than explicitly modeling the graded nature of clinical uncertainty. Inspired by recent uncertainty-aware approaches [83], [84], future work could incorporate explicit uncertainty scores to weight cross-modal alignment, assigning stronger weights to definite findings while attenuating uncertain expressions. Finally, though RIHA's hierarchical alignment framework demonstrates strong theoretical potential for cross-modal adaptation, generalizability is limited due to our current evaluation focusing on chest X-rays. For CT scans, adaptation presents domain-specific challenges including slice interdependency modeling, variable slice thickness, contrast phase dependencies, and specialized preprocessing requirements for CT's wide dynamic range. MRI adaptation introduces complex considerations such as multi-sequence fusion for T1, T2, FLAIR sequences, sequence-specific pathology visibility, and motion artifact handling. While our modular pyramid-based architecture provides flexibility for different report formats, structured templates introduce challenges like template constraint integration for standardized systems and cross-institutional format variations. Despite these modality-specific challenges, our hierarchical alignment principles offer substantial adaptability potential, though practical implementation would require domain-specific modifications and extensive validation. However, accurately capturing rare and subtle findings remains a core challenge in medical report generation. To address this, future work could benefit from the robust in-context learning capabilities of advanced vision-language models to further refine performance.

VI. CONCLUSION

In this work, we propose RIHA, a report-image hierarchical alignment model designed to achieve fine-grained alignment by hierarchically aligning textual and visual information at the paragraph, sentence, and word levels. The optimal transport principle is leveraged for the fine-grained alignment goal, and the relative positional embedding strategy helps enhance the alignment further at the token level. Extensive experiments on benchmark chest X-ray image-report datasets, including IU-Xray and MIMIC-CXR, demonstrate that this approach significantly improves the comprehensiveness and precision of generated radiology reports. These results indicate that a carefully designed fine-grained alignment network can effectively enhance the quality of cross-modal medical report generation.

REFERENCES

[1] K. Konstantinidis, "The shortage of radiographers: A global crisis in healthcare," *Journal of medical imaging and radiation sciences*, vol. 55, no. 4, p. 101333, 2024.

[2] W. Chen, L. Shen, J. Lin, J. Luo, X. Li, and Y. Yuan, "Fine-grained image-text alignment in medical imaging enables explainable cyclic image-report generation," *arXiv preprint arXiv:2312.08078*, 2023.

[3] G. Reale-Nosei, E. Amador-Domínguez, and E. Serrano, "From vision to text: A comprehensive review of natural image captioning in medical diagnosis and radiology report generation," *Medical Image Analysis*, p. 103264, 2024.

[4] I. Hartsock and G. Rasool, "Vision-language models for medical report generation and visual question answering: A review," *Frontiers in Artificial Intelligence*, vol. 7, p. 1430984, 2024.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[6] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," *arXiv preprint arXiv:1711.08195*, 2017.

[7] Z. Li, L. T. Yang, B. Ren, X. Nie, Z. Gao, C. Tan, and S. Z. Li, "Mlip: Enhancing medical visual representation with divergence encoder and knowledge-guided contrastive learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 704–11 714.

[8] M. Chatterjee and A. G. Schwing, "Diverse and coherent paragraph generation from images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 729–744.

[9] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya *et al.*, "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 590–597.

[10] J. Wang, A. Bhalerao, and Y. He, "Cross-modal prototype driven network for radiology report generation," in *European Conference on Computer Vision*. Springer, 2022, pp. 563–579.

[11] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," *arXiv preprint arXiv:2010.16056*, 2020.

[12] J. Wang, A. Bhalerao, T. Yin, S. See, and Y. He, "Camanet: class activation map guided attention network for radiology report generation," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 2199–2210, 2024.

[13] Z. Zhang, Y. Yu, Y. Chen, X. Yang, and S. Y. Yeo, "Medunifier: Unifying vision-and-language pre-training on medical data with vision generation task using discrete visual representations," *arXiv preprint arXiv:2503.01019*, 2025.

[14] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," *arXiv preprint arXiv:2204.13258*, 2022.

[15] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest x-ray report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3334–3343.

[16] Z. Huang, X. Zhang, and S. Zhang, "Kiut: Knowledge-injected u-transformer for radiology report generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 809–19 818.

[17] M. Li, W. Cai, K. Verspoor, S. Pan, X. Liang, and X. Chang, "Cross-modal clinical graph transformer for ophthalmic report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 20 656–20 665.

[18] H. Zhao, J. Chen, L. Huang, T. Yang, W. Ding, and C. Li, "Automatic generation of medical report with knowledge graph," in *Proceedings of the 2021 10th International Conference on Computing and Pattern Recognition*, 2021, pp. 1–1.

[19] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 910–12 917.

[20] S. Yang, X. Wu, S. Ge, Z. Zheng, S. K. Zhou, and L. Xiao, "Radiology report generation with a learned knowledge base and multi-modal alignment," *Medical Image Analysis*, vol. 86, p. 102798, 2023.

[21] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Medical image analysis*, vol. 80, p. 102510, 2022.

[22] Y. Li, B. Yang, X. Cheng, Z. Zhu, H. Li, and Y. Zou, "Unify, align and refine: Multi-level semantic alignment for radiology report generation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 2863–2874.

[23] A. Liu, Y. Guo, J.-h. Yong, and F. Xu, "Multi-grained radiology report generation with sentence-level image-language contrastive learning," *IEEE Transactions on Medical Imaging*, vol. 43, no. 7, pp. 2657–2669, 2024.

[24] S. Y. Yeo, X. Xie, I. Sazonov, and P. Nithiarasu, "Level set segmentation with robust image gradient energy and statistical shape prior," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 3397–3400.

[25] X. Yang, Y. Su, R. Duan, H. Fan, S. Y. Yeo, C. Lim, L. Zhong, and R. S. Tan, "Cardiac image segmentation by random walks with dynamic shape constraint," *IET Computer Vision*, vol. 10, no. 1, pp. 79–86, 2016.

[26] X. Yang, Z. Zeng, S. Y. Yeo, C. Tan, H. L. Tey, and Y. Su, "A novel multi-task deep learning model for skin lesion segmentation and classification," *arXiv preprint arXiv:1703.01025*, 2017.

[27] J. Liu, S. Y. Tan, X. Yang, Y. Xu, and S. Y. Yeo, "Effdnet: A scribble-supervised medical image segmentation method with enhanced foreground feature discrimination," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 194–204.

[28] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[29] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[31] S. Santhanam, "Context based text-generation using lstm networks," *arXiv preprint arXiv:2005.00048*, 2020.

[32] M. Sirshar, M. F. K. Paracha, M. U. Akram, N. S. Alghamdi, S. Z. Y. Zaidi, and T. Fatima, "Attention based automated radiology report generation using cnn and lstm," *Plos one*, vol. 17, no. 1, p. e0262209, 2022.

[33] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7008–7024.

[34] B. Dai, S. Fidler, R. Urtasun, and D. Lin, "Towards diverse and natural image descriptions via a conditional gan," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2970–2979.

[35] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[36] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei *et al.*, "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*. Springer, 2020, pp. 121–137.

[37] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," *arXiv preprint arXiv:2108.10904*, 2021.

[38] X. Hou, Y. Luo, W. Song, Y. Guo, W. You, and S. Li, "Radiographic reports generation via retrieval enhanced cross-modal fusion," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2024, pp. 2032–2039.

[39] Y. Yang, X. You, K. Zhang, Z. Fu, X. Wang, J. Ding, J. Sun, Z. Yu, Q. Huang, W. Han *et al.*, "Spatio-temporal and retrieval-augmented modelling for chest x-ray report generation," *IEEE Transactions on Medical Imaging*, 2025.

[40] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9049–9058.

[41] S. Li, P. Qiao, L. Wang, M. Ning, L. Yuan, Y. Zheng, and J. Chen, "An organ-aware diagnosis framework for radiology report generation,"

IEEE Transactions on Medical Imaging, vol. 43, no. 12, pp. 4253–4265, 2024.

[42] Y. Tian, F. Xia, and Y. Song, "Diffusion networks with task-specific noise control for radiology report generation," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 1771–1780.

[43] Y. Jin, W. Chen, Y. Tian, Y. Song, and C. Yan, "Improving radiology report generation with multi-grained abnormality prediction," *Neurocomputing*, vol. 600, p. 128122, 2024.

[44] F. Wang, Y. Zhou, S. Wang, V. Vardhanabhuti, and L. Yu, "Multi-granularity cross-modal alignment for generalized medical visual representation learning," *Advances in neural information processing systems*, vol. 35, pp. 33 536–33 549, 2022.

[45] T. Zhu, Q. Liu, F. Wang, Z. Tu, and M. Chen, "Unraveling cross-modality knowledge conflicts in large vision-language models," *arXiv preprint arXiv:2410.03659*, 2024.

[46] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "Clip-adapter: Better vision-language models with feature adapters," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 581–595, 2024.

[47] Y. Xia, H. Huang, J. Zhu, and Z. Zhao, "Achieving cross modal generalization with multimodal unified representation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 63 529–63 541, 2023.

[48] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019, 2019, p. 6558.

[49] D. Yang, S. Huang, H. Kuang, Y. Du, and L. Zhang, "Disentangled representation learning for multimodal emotion recognition," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 1642–1651.

[50] Z. Liu, B. Zhou, D. Chu, Y. Sun, and L. Meng, "Modality translation-based multimodal sentiment analysis under uncertain missing modalities," *Information Fusion*, vol. 101, p. 101973, 2024.

[51] J. Tian, K. Wang, X. Xu, Z. Cao, F. Shen, and H. T. Shen, "Multimodal disentanglement variational autoencoders for zero-shot cross-modal retrieval," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022, pp. 960–969.

[52] Y. Zeng, W. Yan, S. Mai, and H. Hu, "Disentanglement translation network for multimodal sentiment analysis," *Information Fusion*, vol. 102, p. 102031, 2024.

[53] Y. Li, Y. Wang, and Z. Cui, "Decoupled multimodal distilling for emotion recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 6631–6640.

[54] M. Li, D. Yang, Y. Lei, S. Wang, S. Wang, L. Su, K. Yang, Y. Wang, M. Sun, and L. Zhang, "A unified self-distillation framework for multimodal sentiment analysis with uncertain missing modalities," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 9, 2024, pp. 10 074–10 082.

[55] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio, "Learning with a wasserstein loss," *Advances in neural information processing systems*, vol. 28, 2015.

[56] L. Yang, J. Liu, S. Hong, Z. Zhang, Z. Huang, Z. Cai, W. Zhang, and B. Cui, "Improving diffusion-based image synthesis with context prediction," *Advances in Neural Information Processing Systems*, vol. 36, pp. 37 636–37 656, 2023.

[57] F. L. Hitchcock, "The distribution of a product from several sources to numerous localities," *Journal of mathematics and physics*, vol. 20, no. 1-4, pp. 224–230, 1941.

[58] A. R. Lahitani, A. E. Permanasari, and N. A. Setiawan, "Cosine similarity to determine similarity measure: Study case in online essay assessment," in *2016 4th International conference on cyber and IT service management*. IEEE, 2016, pp. 1–6.

[59] S. Kullback, "Kullback-leibler divergence," 1951.

[60] J.-D. Rolle, "Various issues around the l1-norm distance," *arXiv preprint arXiv:2110.04787*, 2021.

[61] I. Dokmanic, R. Parhizkar, J. Ranieri, and M. Vetterli, "Euclidean distance matrices: essential theory, algorithms, and applications," *IEEE Signal Processing Magazine*, vol. 32, no. 6, pp. 12–30, 2015.

[62] M. Cuturi, "Lightspeed computation of optimal transportation distances," *Advances in Neural Information Processing Systems*, vol. 26, no. 2, pp. 2292–2300, 2013.

[63] H. Fan, H. Su, and L. J. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 605–613.

[64] R. Jonker and T. Volgenant, "A shortest augmenting path algorithm for dense and sparse linear assignment problems," in *DGOR/NSOR: Papers of the 16th Annual Meeting of DGOR in Cooperation with NSOR/Vorträge der 16. Jahrestagung der DGOR zusammen mit der NSOR*. Springer, 1988, pp. 622–622.

[65] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[66] C. Villani *et al.*, *Optimal transport: old and new*. Springer, 2008, vol. 338.

[67] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[68] K. Wu, H. Peng, M. Chen, J. Fu, and H. Chao, "Rethinking and improving relative position encoding for vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 033–10 041.

[69] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[70] Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," *Advances in neural information processing systems*, vol. 31, 2018.

[71] F. Liu, X. Wu, S. Ge, W. Fan, and Y. 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*, 2021, pp. 13 753–13 762.

[72] A. E. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[73] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[74] M. Denkowski and A. Lavie, "Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 85–91.

[75] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[76] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[77] F. Liu, S. Ge, Y. Zou, and X. Wu, "Competence-based multimodal curriculum learning for medical report generation," *arXiv preprint arXiv:2206.14579*, 2022.

[78] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and explainable region-guided radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7433–7442.

[79] H. Jin, H. Che, Y. Lin, and H. Chen, "Promptmrg: Diagnosis-driven prompts for medical report generation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2607–2615.

[80] T. Gu, D. Liu, Z. Li, and W. Cai, "Complex organ mask guided radiology report generation," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2024, pp. 7995–8004.

[81] S. Bu, T. Li, Y. Yang, and Z. Dai, "Instance-level expert knowledge and aggregate discriminative attention for radiology report generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 194–14 204.

[82] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.

[83] Z. Wang, S. Yan, K. Yin, X. Zhang, and W. K. Cheung, "Curv: Coherent uncertainty-aware reasoning in vision-language models for x-ray report generation," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[84] I. Najdenkoska, X. Zhen, M. Worring, and L. Shao, "Uncertainty-aware report generation for chest x-rays by variational topic inference," *Medical Image Analysis*, vol. 82, p. 102603, 2022.