arXiv:2505.03401v2 [cs.CV] 29 Aug 2025

DDaTR: Dynamic Difference-aware Temporal Residual Network for Longitudinal Radiology Report Generation

Shanshan Song, Hui Tang, Honglong Yang, Xiaomeng Li *Member, IEEE*

***Abstract*— Radiology Report Generation (RRG) automates the creation of radiology reports from medical imaging, enhancing the efficiency of the reporting process. Longitudinal Radiology Report Generation (LRRG) extends RRG by incorporating the ability to compare current and prior exams, facilitating the tracking of temporal changes in clinical findings. Existing LRRG approaches only extract features from prior and current images using a visual pre-trained encoder, which are then concatenated to generate the final report. However, these methods struggle to effectively capture both spatial and temporal correlations during the feature extraction process. Consequently, the extracted features inadequately capture the information of difference across exams and thus underrepresent the expected progressions, leading to sub-optimal performance in LRRG. To address this, we develop a novel dynamic difference-aware temporal residual network (DDaTR). In DDaTR, we introduce two modules at each stage of the visual encoder to capture multi-level spatial correlations. The Dynamic Feature Alignment Module (DFAM) is designed to align prior features across modalities for the integrity of prior clinical information. Prompted by the enriched prior features, the dynamic difference-aware module (DDAM) captures favorable difference information by identifying relationships across exams. Furthermore, our DDaTR employs the dynamic residual network to unidirectionally transmit longitudinal information, effectively modeling temporal correlations. Extensive experiments demonstrated superior performance over existing methods on three benchmarks, proving its efficacy in both RRG and LRRG tasks. Our code is published at** `https://github.com/xmed-lab/DDaTR`

***Index Terms*— Radiology report generation, Longitudinal radiology report generation, Dynamic difference-awareness, Longitudinal multimodal encoder.**

I. Introduction

Radiology Report Generation (RRG) focuses on the automated creation of coherent and precise textual descriptions for radiology images, thereby reducing radiologists' workload [1], [2]. In a typical radiology workflow, radiologists not only deliver comprehensive analyses of clinical findings from the current exam, but also compare the current images with prior images and reports to identify significant progressions in key observations, assess treatment response, or detect early signs of disease recurrence. To accurately generate the report with clinical progressions, Longitudinal Radiology Report Generation (LRRG) focuses on integrating a patient's current and prior exams to provide a comprehensive description of their condition in the report [3]. Compared to RRG, this task monitors the progression of a patient's condition and can more effectively assist doctors in developing subsequent treatment plans [3], [4], [6]. Moreover, by leveraging historical clinical data, LRRG promotes diagnostic consistency across different visits and among different practitioners, reducing variability in interpretations. However, it is more challenging as it requires modeling spatio-temporal information to capture the core progression, and there is limited research addressing this problem. Consequently, our research is dedicated to addressing the challenges associated with the LRRG task.

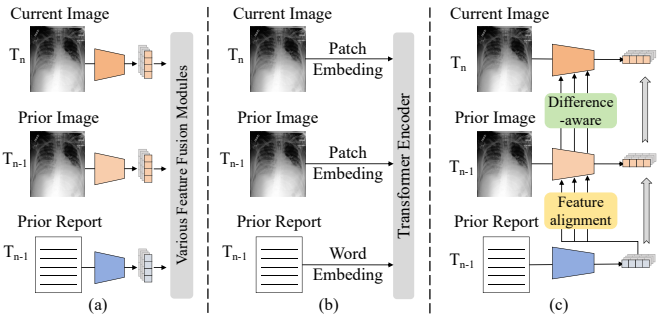


Fig. 1. Comparison of our method with previous LRRG approaches. **(a)** represents methods that encode different modalities separately and subsequently fuse them using various feature fusion modules [3]–[5]. **(b)** denotes methods that embed longitudinal information into tokens, which are then processed collectively through a transformer-based encoder [6]–[8]. **(c)** is our proposed DDaTR, a novel approach for longitudinal information fusion that leverages prior image-text feature alignment and longitudinal difference-aware information.

Previous RRG works primarily adopt the encoder-decoder structure from the image captioning field [9], [10]. While image captioning typically targets short text descriptions, RRG requires the creation of fine-grained, long-text paragraphs for clinical findings. To achieve it, some studies improved the encoder-decoder framework by designing new networks and modules [11]–[14]. Some methods incorporated medical knowledge graphs or extra medical knowledge [15]–[18]. Recently, several approaches have employed multi-task learning

Corresponding author: Xiaomeng Li

This work was supported by a research grant from the Joint Research Scheme (JRS) under the National Natural Science Foundation of China (NSFC) and the Research Grants Council (RGC) of Hong Kong (Project No. N_HKUST654/24), as well as a grant from the RGC of the Hong Kong Special Administrative Region, China (Project No. R6005-24).
Shanshan Song, Hui Tang, Honglong Yang, and Xiaomeng Li are with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong, SAR, China (email: ssongan@connect.ust.hk; eehtang@ust.hk; hyangdh@connect.ust.hk; eexmli@ust.hk)

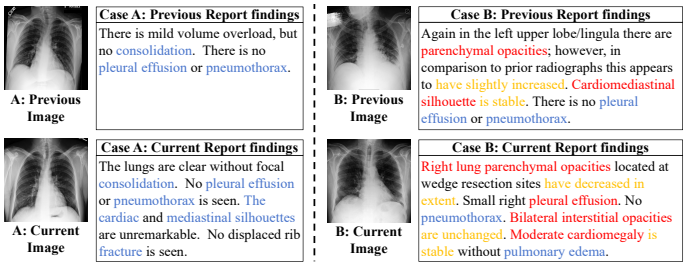


Fig. 2. Different LRRG task examples. **Blue** text indicates 'negative observations', **red** text indicates 'positive observations', and **yellow** text denotes the 'progressions'. Both case A and case B include prior image and report for reference. However, in case A, there is no valuable longitudinal information to describe in the current report findings, indicating a minimal contribution from prior information. Conversely, case B primarily describes longitudinal information, demonstrating a significant contribution from prior information.

that facilitate the core generation task with other vision tasks, e.g., object detection to extract abnormal regions and anatomical structures [19], classification to predict observation labels [20], [21]. Multi-task learning has proven effective in capturing fine-grained information and enhancing interpretability. However, above RRG approaches fail to generate longitudinal descriptions that align accurately with radiologists' intentions.

To address the challenge of LRRG, several approaches have been proposed to fuse the multi-period and multimodal features before generating the final reports. As shown in Fig. 1 (a), one category of works [3]–[5] applies various feature fusion modules, such as cross-attention modules, after encoding each image or text with their respective pre-training encoders, and then feeds the fused features into a decoder to generate the corresponding reports. As shown in Fig. 1 (b), another category [6]–[8] leverages the powerful encoding capabilities of transformers to embed multi-period inputs, which uses modality-type and temporal encoding to differentiate tokens from different modalities and periods before decoding the report. Moreover, RECAP [8] uses extra labels from the Chest ImaGenome dataset [22] to help learn the progressions across different periods during the encoding process.

However, the direct application of general visual pretraining encoders to the LRRG task for longitudinal feature extraction presents two key limitations, which we summarize as follows: (1) **Inaccurate perception of temporal correlations.** Existing methods overlook the different contributions of prior exams to current reports. As an example shown in Fig. 2, even when prior images and reports are available, it is not always necessary to include longitudinal information in the current report findings. When prior exam is unavailable, the model should rely solely on current-phase features. Moreover, the unidirectional nature of temporal information transmission only from prior to current is also neglected. When extracting visual features from the prior image, the model should consider the prior report only, without incorporating any information from the current examination. (2) **Inaccurate perception of spatial correlations.** In this work, spatial correlations refer to the relationships among local-level features within a single X-ray image, as well as the correspondences of detailed textures and structural patterns across a longitudinal image pair. For single image, previous methods directly extract prior image features without aligning them with the corresponding reports, which may contain clinical prior knowledge. Due to this misalignment, the extracted visual features fail to accurately perceive known observations or abnormalities. Additionally, traditional fusion methods independently encode each image and then fuse them afterward, disregarding the multi-level spatial relationships between longitudinal images during the encoding process which particularly neglects the relationship of low-level details and textures.

To address the above limitations, we design a novel dynamic difference-aware temporal residual network (DDaTR). Addressing limitation (1), our entire network dynamically integrates temporal information from prior image features at different levels of abstraction into current image features. The residual connections are used to unidirectionally transmit temporal information, which is more suitable for modeling longitudinal image-text relationships by adaptively learning the contributions of different prior features. Furthermore, this unidirectional information flow aligns better with clinical practice, enhancing the focus on generating the current report. Addressing limitation (2), we introduce two modules at each stage of the visual encoder: the Dynamic Feature Alignment Module (DFAM) and the Dynamic Difference-aware Module (DDAM). DFAM employs the entire prior report's features to guide the prior image encoding, thereby better aligning prior image-text features. The DDAM leverages differences between two image features to adaptively enhance the current image feature, effectively addressing the insufficient perception of feature differences across multi-period images.

In summary, the contributions of our work are as follows:

- We propose a novel dynamic difference-aware temporal residual network (DDaTR) to better encode longitudinal information by adaptively learning the multi-level abstractions of different prior features and modeling the unidirectional information flow.
- We introduce two plug-and-play modules: DFAM and DDAM. These modules can be integrated into any multi-scale visual encoder for longitudinal feature extraction.
- We conduct assessments not only on the standard MIMIC-CXR benchmark [23] but also under various settings with different proportions of longitudinal reports, demonstrating the robustness and versatility of our approach across diverse clinical scenarios. Extensive experiments and analyses indicate that our method aligns more closely with clinical applications and achieves state-of-the-art results in both RRG and LRRG tasks.

II. Related Work

A. Radiology Report Generation

Recently, automated Radiology Report Generation (RRG) has attracted significant research interest. Most RRG studies use CNN-based or Transformer-based visual pre-training backbones to extract visual features and RNN-based [24]–[26] or Transformer-based structures [11], [27], [28] to decode the corresponding reports. RRG primarily focuses on the characteristics of radiographic images and reports which present more challenges compared to image captioning tasks.

To generate long reports with detailed abnormal findings and enhanced cross-modal alignment, several studies have focused on improving the encoder-decoder framework by designing new networks and incorporating various modules. For instance, R2Gen [11], R2GenCMN [12], CMM-RL [13] and METransformer [14]. Although these methods design new networks to generate more accurate and coherent radiology reports, the end-to-end image-to-text generation framework lacks fine-grained clinical guidance and explicit domain knowledge. This limitation hinders the clinical accuracy of the generated reports, resulting in a low Clinical Efficacy (CE) score. Considering the medical domain knowledge and severe visual and textual bias in RRG, several works leverage medical knowledge graphs or extra medical prior knowledge to aid in generating accurate radiology reports, such as KGAE [15], KMGen [16], KiUT [17] and DCL [18]. However, constructing high-quality knowledge graphs is both costly and labor-intensive. Furthermore, the above methods that rely on specific prior knowledge often fail to generalize effectively across new modalities or datasets.

In RRG, accurately describing the presence or absence of clinical observations is more crucial than the linguistic style or expression used. To capture more fine-grained clinical features and enhance clinical efficacy, many approaches employ multi-task learning methods. These methods not only support the core report generation task but also integrate other vision-related tasks to aid the accurate report generation. RGRG [19] utilizes the object detection task to extract abnormal regions and anatomical structures from the image, and then feeds them to a final report decoder. ORGAN [20] leverages the classification task to predict observation labels and attributes of diseases from images, and then uses these information to enhance the decoding process. However, these methods employ fine-grained information in an implicit way, facing challenges in effectively establishing a correspondence between intermediate features and the report. To overcome this, PromptMRG [21] uses an extra disease classification branch, and then generate final report with the guidance of diagnosis-aware prompts. Due to make use of diagnostic prompts to explicitly guide the generation process, PromptMRG has better explainability and meets the actual clinical workflow. Therefore, we choose the framework with diagnosis-aware prompts as our strong baseline for report generation.

B. Longitudinal Radiology Report Generation

Previous RRG methods that focus on single-period data overlook historical records, which can result in inaccuracies and hallucinations about unobserved priors when training language models for report generation [29]. Recently, to improve the generation of reports that accurately reflect clinical progressions, various approaches have been developed to integrate multi-period and multimodal features, utilizing longitudinal information to produce more accurate final reports.

One category of works utilizes various feature fusion modules, such as cross-attention modules and projection modules, to integrate multi-period visual and textual features after encoding each image or text with their respective pre-training encoders, and then feeds the fused features into a decoder to generate the corresponding reports. In this study [4], Faster R-CNN is employed to extract anatomical regions from both current and prior images. Subsequently, a longitudinal projection module is utilized to align, concatenate, and project these region representations into a cohesive joint representation. LR2Gen [3] captures information from longitudinal patient visit records using a cross-attention-based multimodal fusion module and a hierarchical memory-driven decoder. HERGen [5] employs a group causal transformer to efficiently integrate longitudinal data across patient visits.

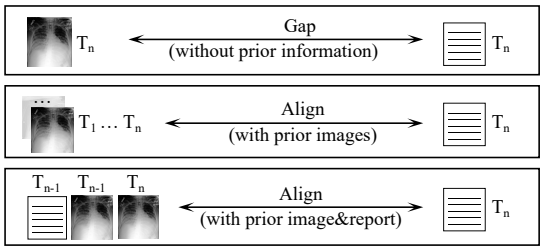


Fig. 3. Characteristics of Longitudinal Report Generation. The top panel highlights the gap between the current image and the corresponding report due to the lack of prior information. The middle panel illustrates how the final radiology report is constructed by integrating images from $\boldsymbol{T_1}$ to $\boldsymbol{T_n}$, taking into account both the patient's current condition and their historical progressions in the real-world. The bottom panel shows the longitudinal information from $\boldsymbol{T_{n-1}}$ report, $\boldsymbol{T_{n-1}}$ and $\boldsymbol{T_n}$ images are also aligned with the target $\boldsymbol{T_n}$ report which is used in our work.

Another category leverages the powerful encoding capabilities of transformers to embed multi-period images and texts. BioViL-T [6] utilizes a CNN–Transformer hybrid multi-image encoder to extract longitudinal visual features. CXRMate [7] incorporates longitudinal information by using prior reports as prompts, which are fed into a transformer decoder. Additionally, RECAP [8] utilizes extra progression labels to assist the visual transformer-based encoder in learning the progressions across different periods.

Although adding longitudinal features could alleviate the hallucinations about unobserved priors when training language models for report generation, there are still limitations in capturing the important longitudinal correlations. Firstly, they overlook the different contributions of previous image content (e.g., different viewpoints, limited longitudinal information or varying time intervals between examinations). This limitation leads to insufficient global temporal correlations. Secondly, previous feature fusion is used after visual encoder. Specifically, they use general visual pre-training encoders (e.g.,ResNet [30], SwinTransformer [31]) to extract visual features of each images independently, and then they leverage cross-attention based modules or transformer blocks to encode longitudinal features. However, the encoding process does not consider the relationships between multi-level features of longitudinal images, especially the low-level features including important details and texture information. This limitation leads to inaccurate local temporal correlations. Thirdly, previous methods fail to account for the prior knowledge in previous report during the visual feature extraction. This report documents the patient's clinical findings and diagnoses, reflecting the underlying clinical intent of the examination, which is important for perceiving prior visual semantics.

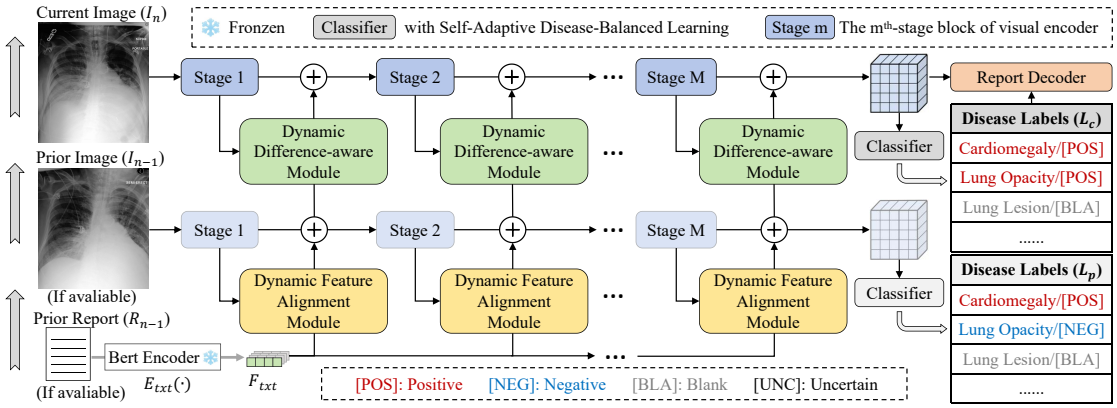


Fig. 4. The framework of our proposed DDaTR. First, the prior report is processed through the text encoder $\boldsymbol{E_{txt}}$ to extract textual features $\boldsymbol{F_{txt}}$. Next, the current image, prior image, and $\boldsymbol{F_{txt}}$ from the prior report are input into the DDaTR which incorporates the Dynamic Difference-aware Module and Dynamic Feature Alignment Module at multi-level encoding stage. The output features for the prior image and the current image are separately fed into different classifiers to obtain the respective disease labels. Finally, the current disease labels are used as prompts and input into the decoder, where the current image features are interactively fused with the decoder via cross-attention, ultimately generating the current report.

III. Proposed Method

To address the challenges of the LRRG task, previous approaches primarily adopted two types of settings. The first setting focused exclusively on generating longitudinal reports [3], [4], utilizing datasets that only included cases with prior records and disregarding data with only a single exam record. The second setting considered both single-period and multi-period data [6], [8]. Their methodologies aimed to enhance the overall performance of general radiology report generation tasks, rather than focusing solely on longitudinal reports.

We believe that the incorporation of prior information is intended to improve the quality of the final report generation. If it proves effective only on longitudinal data but fails to surpass the performance of single-period report generation methods, the significance of incorporating prior data is limited. Therefore, we opt for the second setting. Our goal is to refine the utilization of multi-period information so that the performance of report generation is not only robust on multi-period datasets but also achieves state-of-the-art results in both single-period and multi-period report generation tasks.

A. Problem Definition and Our Baseline Framework

Problem Definition. We follow the setting in Recap [8], which involves generating a current report based on both the current image and, when available, prior image and report. However, our approach differs from Recap, which relies on the Chest ImaGenome dataset [22] that selects only Anteroposterior (AP) or Posteroanterior (PA) scans as priors and excludes lateral views. We obtain the longitudinal scans of each patient by chronologically organizing different studies according to corresponding timestamps. Our setting includes consideration of lateral views as well. As shown in Fig.3, the work in this paper is based on the hypothesis that longitudinal information from the T_{n-1} report, T_{n-1} image and T_n image is aligned with the target T_n report.

Specifically, define the current exam as the n^{th} exam. The input consists of the current image $I_n \in \mathbb{I}^{3\times224\times224}$, the prior image $I_{n-1} \in \mathbb{I}^{3\times224\times224}$, and the prior report $R_{n-1} = \{r_1, r_2, ...r_{l_{n-1}}\}$. The output is the current report $R_n = \{r_1, r_2, ...r_{l_n}\}$. l_{n-1} and l_n are the word length of prior and current report respectively. If $n = 1$, indicating that this is the patient's first exam, the I_{n-1} and R_{n-1} is None.

Our Baseline Framework. The generation framework presented in our paper is based on PromptMRG [21], which integrates an additional classification branch for generating diagnostic prompts used in producing the final report. In the disease classification branch, classification labels are derived using CheXbert [32], which converts reports into 14 predefined diseases. Each disease is described by one of four attributes: blank, positive, negative, or uncertain. Specifically, the encoder $E_{img}(\cdot)$ is firstly used to extract the visual feature F_{img} from the input image I_n.

$$F_{img} = E_{img}\left(I_n\right) \tag{1}$$

Then, the visual feature F_{img} is passed through a classifier $f_{cls}(\cdot)$ with self-adaptive disease-balanced learning (SDL) [21], which outputs 14 predefined disease labels $L_{label} \in \mathbb{R}^{14}$.

$$L_{label} = f_{cls}\left(F_{img}\right) \tag{2}$$

Subsequently, these disease labels are converted into corresponding prompt tokens T_{label}, with each token corresponding to a specific disease. For each disease, one of the four new tokens, including [BLA], [POS], [NEG], and [UNC], is utilized to represent the disease attribute. Finally, both the prompt tokens T_{label} and the visual feature F_{img} are fed into a decoder $D(\cdot)$ to generate the target report R_n.

$$R_n = D\left(T_{label}, F_{img}\right) \tag{3}$$

To perform LRRG based on the above baseline, we propose Dynamic Difference-aware Temporal Residual Network (DDaTR) $E_{ddatr}(\cdot)$ to integrate longitudinal multimodal information to get better visual features. Therefore, the encoding process in our paper is formalized as follows:

$$F_{img} = \begin{cases} E_{ddatr}\left(I_n\right) & \text{if } n = 1 \\ E_{ddatr}\left(I_n, I_{n-1}, R_{n-1}\right), & \text{if } n > 1 \end{cases} \tag{4}$$

where F_{img} represents the output longitudinal visual feature, which differs from the baseline visual feature defined in Eqn. (1). After obtaining F_{img}, the subsequent step is the same as described in Eqn. (2) and Eqn. (3).

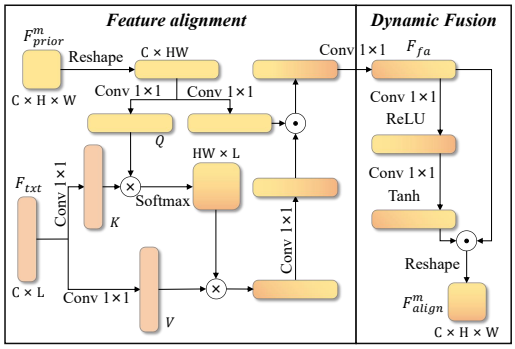


Fig. 5. The structure of Dynamic Feature Alignment Module (DFAM). Given prior image features $\boldsymbol{F^m_{prior}}$ and prior text features $\boldsymbol{F_{txt}}$ as inputs, the module processes these through feature alignment and dynamic fusion sequentially. Finally, the output is visual features $\boldsymbol{F^m_{align}}$ that are aligned with the text and dynamically fused.

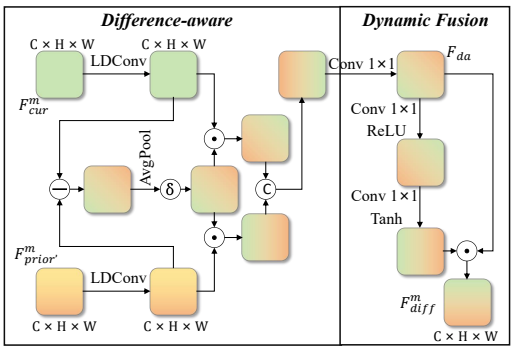


Fig. 6. The structure of Dynamic Difference-aware Module (DDAM). Given prior aligned visual features $\boldsymbol{\hat{F}^m_{prior}}$ and current visual features $\boldsymbol{F^m_{cur}}$ as inputs, the module processes these through difference-aware and dynamic fusion sequentially. Finally, the output is visual features $\boldsymbol{F^m_{diff}}$ that are compared with the prior feature and dynamically fused.

B. Overview of Our Proposed DDaTR

To better extract features from multimodal longitudinal inputs and effectively model the temporal relationships, we propose the Dynamic Difference-aware Temporal Residual Network (DDaTR). As shown in Fig. 4, our framework employs the encoder-decoder architecture with a disease classification branch. In this framework, the encoder $E_{ddatr}(\cdot)$ extracts the longitudinal features F_{img} from the current image I_n, the prior image I_{n-1}, and the prior report R_{n-1}. Specifically, we begin by encoding the prior report R_{n-1} using the pretrained BERT [33] model $E_{txt}(\cdot)$ to extract the text feature F_{txt}.

$$F_{txt} = E_{txt}\left(R_{n-1}\right) \tag{5}$$

During the training process, the parameters of $E_{txt}(\cdot)$ are frozen without gradient updates. This not only reduces the number of training parameters but also focuses model optimization on the visual features. The text feature F_{txt} is employed to align with the prior image features at each stage of the visual encoder, thereby enhancing the learned visual features of I_{n-1} and facilitating the following perception of differences between two-period images.

Firstly, the output prior visual feature from the m^{th} stage of the visual encoder is denoted as F^m_{prior}. $m \in [1, ..., M]$, where M is the number of stages of the visual encoder:

$$F^m_{prior} = \begin{cases} E^m_{prior}\left(I_{n-1}\right) & \text{if } m = 1 \\ E^m_{prior}\left(\hat{F}^{m-1}_{prior}\right) & \text{if } m > 1 \end{cases} \tag{6}$$

where $\hat{F}^{m-1}_{prior}$ represents the final output feature from the $m-1$ stage and E^m_{prior} denotes the m^{th} stage of prior visual encoder. For each stage of the prior visual encoder, we introduce the Dynamic Feature Alignment Module (DFAM) to dynamically fuse the image feature F^m_{prior} and text feature F_{txt}, resulting in the aligned output feature F^m_{align}.

$$F^m_{align} = DFAM\left(F^m_{prior}, F_{txt}\right) \tag{7}$$

where the mechanics of the $DFAM(\cdot)$ are detailed in Sec. III-C. Next, this aligned output is added to F^m_{prior} via a residual connection to obtain the final prior visual feature. The formula is given by:

$$\hat{F}^m_{prior} = F^m_{prior} + F^m_{align} \tag{8}$$

where $\hat{F}^m_{prior}$ represents the final prior visual feature of the prior m^{th} stage, which is subsequently fed into the next stage.

Secondly, the output current visual feature from the m^{th} stage of current visual encoder is denoted as F^m_{cur} as follows:

$$F^m_{cur} = \begin{cases} E^m_{cur}\left(I_n\right) & \text{if } m = 1 \\ E^m_{cur}\left(\hat{F}^{m-1}_{cur}\right) & \text{if } m > 1 \end{cases} \tag{9}$$

where E^m_{cur} denotes the m^{th} stage of current visual encoder and $\hat{F}^{m-1}_{cur}$ represents the final output feature from the current stage. Utilizing our proposed Dynamic Difference-aware Module (DDAM), we incorporate F^m_{cur} with $\hat{F}^m_{prior}$, resulting in the difference-aware output F^m_{diff}.

$$F^m_{diff} = DDAM\left(F^m_{cur}, \hat{F}^m_{prior}\right) \tag{10}$$

The mechanics of the $DDAM(\cdot)$ are detailed in Sec. III-D. This output is also added to F^m_{cur} using a residual connection, subsequently feeding into the next stage. Ultimately, this process culminates in the acquisition of the current visual feature $\hat{F}^m_{cur}$. The equation can be expressed as:

$$\hat{F}^m_{cur} = F^m_{cur} + F^m_{diff} \tag{11}$$

During training, the final output prior and current visual features of the last stage deliver to two different classifiers to get their corresponding observation label: L_p, L_c. Then the decoder D is utilized to generate the current report R_n, conditioned on both the longitudinal visual features $\hat{F}^M_{cur}$ and the diagnosis-driven prompts derived from L_p.

$$R_n = D\left(L_c, \hat{F}^M_{cur}\right) \tag{12}$$

It is worth noting that the classifier for the prior is used as a regularizer to assist in learning the prior-aligned visual feature, which does not need to be sent to the decoder. This is because the classifier for the prior part intended to convey additional supervision signals from prior image-text pairs which are groundtruth, effectively acting as a regularizer to ease the training process. Specifically, (1) it guides the learning of prior-aligned visual features by supervising the extraction of prior visual features using prior observation labels. This encourages the model to learn disease-discriminative patterns from prior visits, enabling it to extract relevant and useful prior information when generating reports from current images. (2)

due to the current features are supervised using the current labels, this design ensures semantic consistency between the prior and current visual encoders, which supports effective feature alignment and fusion across stages. As a result, our framework is better able to model longitudinal dependencies, leading to more accurate and context-aware report generation.

C. Dynamic Feature Alignment Module

Prior report typically contain valuable clinical prior knowledge that goes beyond what is directly observable in the image, including the clinician's focus and key diagnostic observations. Therefore, our goal is not merely to extract visual features from the prior image, but to enhance them by aligning with the semantic information in the report. Therefore, we propose DFAM, which enables the model to inject clinical semantics from the report into the pixel-level representation of the image, allowing the extracted prior features to better reflect clinical meaning. Our DFAM is inspired by the findings of the LAVT [39], which demonstrated that early fusion of image-text features within the intermediate layers of a vision encoder can more effectively align visual features with text. During the prior feature encoding process, we use the prior observation classification task as a supervisory signal. The structure of the DFAM is illustrated Fig. 5, which consists of two components: feature alignment and dynamic fusion.

At the first step, the input features from the two modalities—visual features $F^m_{prior} \in \mathbb{R}^{C\times H\times W}$ at m^{th} stage and textual features $F_{txt} \in \mathbb{R}^{C\times L}$—are aligned. Specifically, F^m_{prior} is mapped as queries, and F_{txt} is mapped as keys and values to conduct scaled dot-product attention [40]. This design choice is made to ensure that the fused representation remains primarily image-driven and facilitates the subsequent perception of visual difference. The textual features are incorporated to inject clinical semantic of reports into the visual representations, thereby promoting better alignment between the prior image features and the underlying clinical intent. We define P as the operation of projection. This process is represented by the following equations:

$$Q = P^q(\text{Reshape}(F^m_{prior})); K = P^k(F_{txt}); V = P^v(F_{txt}) \quad (13)$$

$$F_{att} = P^t(\text{Softmax}\left(\frac{Q^T K}{\sqrt{C}}\right)V^T) \quad (14)$$

where P^k and P^v are 1×1 convolution. P^q and P^t are the 1×1 convolution followed by instance normalization. F_{att} represents the output features after this process. This output $F_{att} \in \mathbb{R}^{HW\times C}$ has the same spatial resolution as the original visual features but integrates semantic information from the text at each image pixel. This design is aligned with our goal of injecting report-level semantic guidance into the visual representation, allowing each spatial position in the image to be enriched with contextually relevant textual semantics. Next, F_{att} are further integrated with the feature mapped from original image features F^m_{prior} as follows:

$$F_{fa} = P^f(P^a(\text{Reshape}(F^m_{prior})) \odot F_{att}) \quad (15)$$

where P^a and P^f are 1×1 convolution followed by ReLU, $\odot$ denotes element-wise multiplication and $F_{fa} \in \mathbb{R}^{HW\times C}$ represents the output of the first feature alignment part.

Secondly, the F_{fa} is fused with the original input image feature F^m_{prior} using a gating mechanism. Specifically, for the aligned features F_{fa}, the process has two blocks B_1 and B_2. The output features F_{dy} are then combined with F^m_{prior} through element-wise multiplication, ultimately producing the final output features F^m_{align} of the DFAM at m^{th} stage of prior encoder. This process is described as follows:

$$F^m_{align} = B_2(B_1(F_{fa})) \odot F_{fa} \quad (16)$$

where $B_2(B_1(F_{fa}))$ denotes the output features of gating mechanism. B_1 is the operation of 1x1 convolution followed by ReLU. B_2 is the operation of 1x1 convolution with Tanh.

D. Dynamic Difference-aware Module

To more effectively capture the multi-level progress information between prior and current images, we propose the Dynamic Difference-aware Module (DDAM). Similar to the DFAM, the DDAM is also integrated at every stage of the visual encoder to enable perception at both low-level and high-level features. Furthermore, this module can adaptively enhance the texture details of both visual features. The structure of the DDAM is illustrated in Fig. 6 and comprises two components: difference awareness and dynamic fusion.

In the first part, we employ the enhanced prior visual features $\hat{F}^m_{prior}$ processed using DFAM. Inspired by Fusion-mamba [41], the Learnable Descriptive Convolution (LDConv) [42] improves texture processing on input feature maps by incorporating learnable descriptors into vanilla convolution operations. Therefore, we also incorporate the LDConv in our DDAM to improve texture detail features. "the texture details" refer to fine-grained visual patterns in the images, such as edges, contours, and subtle structural variations. Specifically, given the current visual feature F^m_{cur} and the prior enhanced visual feature $\hat{F}^m_{prior}$, we first apply two LDConv modules to enhance each feature individually, resulting in the enhanced features F_{enc} and F_{enp}:

$$F_{enc} = LDC_c(F^m_{cur}); F_{enp} = LDC_p(\hat{F}^m_{prior}) \quad (17)$$

Next, we compute the pixel differences between the enhanced features F_{enc} and F_{enp}:

$$F_{pd} = \text{Sigmoid}(\text{AvgPool}(F_{enc} - F_{enp})) \quad (18)$$

where F_{pd} is the output feature of pixel differences. Then, the difference feature F_{pd} is applied to the two enhanced features F_{enc} and F_{enp} to amplify the distinctions between them, obtaining the current difference-aware feature $F_{enc}\odot F_{pd}$ and the prior one $F_{enp}\odot F_{pd}$. Finally, we concatenate them to obtain the difference-aware feature. The process is as follows:

$$F_{da} = P^d(Cat(F_{enc}\odot F_{pd}, F_{enp}\odot F_{pd})) \quad (19)$$

where $Cat(\cdot)$ represents the operation of concatenation. P^d is a 1×1 convolution followed by ReLU activation.

Secondly, the dynamic fusion part of the DDAM is the same as DFAM, designed to ensure that the difference-aware features do not overwhelm the current visual features while allowing a controlled flow of longitudinal information into the next stage. This module dynamically integrates prior information, allowing the model to adaptively select the most relevant features. Specifically, F^m_{cur} represents the visual features, F_{da} means the difference-aware visual features including longitudinal features. When we fuse the F_{cur} and F_{da}, we dynamically add F_{da} to F_{cur} using a gating mechanism. f_{dy} represents

TABLE I

COMPARISON WITH EXISTING REPORT GENERATION METHODS ON MIMIC-CXR, LONGITUDINAL-MIMIC AND IU-XRAY. * INDICATES RESULTS EVALUATED BY US. FOR THE MIMIC-CXR AND IU-XRAY, WE ADOPT THE EVALUATION OUTLINED BY RECAP [8], USING THE MACRO SCORE FOR CE METRICS. FOR THE LONGITUDINAL-MIMIC, WE FOLLOW L-R2GEN [3], CALCULATING THE MICRO SCORE FOR CE METRICS.

Datasets	Methods	Year	Clinical Efficacy			Natural Language Generation			
			Precision	Recall	F1	BLEU-1	BLEU-4	METEOR	ROUGE-L
MIMIC-CXR	R2Gen [11]	2020	0.333	0.273	0.276	0.353	0.103	0.142	0.270
	R2GenCMN [12]	2021	0.344	0.275	0.278	0.353	0.106	0.142	0.278
	KGAE [15]	2021	0.389	0.362	0.355	0.369	0.118	0.153	0.295
	CMM-RL [13]	2022	0.342	0.294	0.292	0.381	0.109	0.151	0.287
	KMGen [16]	2022	0.458	0.348	0.371	0.363	0.115	-	0.284
	KiUT [17]	2023	0.371	0.318	0.321	0.393	0.113	0.160	0.285
	DCL [18]	2023	**0.471**	0.352	0.373	-	0.109	0.150	0.284
	METrans. [14]	2023	0.364	0.309	0.311	0.386	0.124	0.152	0.291
	ORGAN [20]	2023	0.416	0.418	0.385	0.407	0.123	0.162	0.293
	R2GenGPT [34]	2023	0.392	0.387	0.389	0.411	**0.134**	0.160	**0.297**
	RECAP [8]	2023	0.389	0.443	0.393	**0.429**	0.125	**0.168**	0.288
	HERGen [5]	2024	0.415	0.301	0.317	0.395	0.122	0.156	0.285
	PromptMRG* [21]	2024	0.454	0.370	0.389	0.398	0.112	0.157	0.268
	DDaTR(Ours)	2024	0.438	**0.465**	**0.441 (+0.048)**	0.401	0.113	0.162	0.272
Longitudinal-MIMIC	AoANet [35]	2021	0.437	0.249	0.371	0.272	0.080	0.115	0.249
	R2Gen [11]	2020	0.500	0.305	0.379	0.302	0.087	0.124	0.259
	R2GenCMN [12]	2021	0.521	0.396	0.449	0.305	0.085	0.126	0.265
	CNN+Trans [36]	2021	0.445	0.258	0.326	0.299	0.088	0.120	0.263
	LR2Gen [3]	2023	0.538	0.434	0.480	0.343	0.099	0.137	**0.271**
	PromptMRG* [21]	2024	0.502	0.543	0.492	0.390	0.102	0.152	0.263
	DDaTR(Ours)	2024	**0.539**	**0.584**	**0.527 (+0.035)**	**0.396**	**0.105**	**0.156**	0.266
IU-Xray	R2Gen [11]	2020	0.161	0.104	0.071	0.325	0.059	0.131	0.253
	CVT2Dis. [37]	2023	0.325	0.166	0.155	0.383	0.082	0.147	0.277
	M2KT [38]	2023	0.224	0.179	0.151	0.371	0.078	0.153	0.261
	DCL [18]	2023	0.257	0.202	0.177	0.354	0.074	0.152	0.267
	RGRG [19]	2023	0.241	0.224	0.187	0.266	0.063	0.146	0.180
	PromptMRG [21]	2024	0.275	0.288	0.246	0.401	0.098	**0.160**	0.281
	DDaTR(Ours)	2024	**0.304**	**0.291**	**0.262(+0.016)**	**0.421**	**0.103**	**0.160**	**0.307**

our gating module, and the fusion output is $F_{dy} = f_{dy}(F_{da})$, where F_{dy} represents the output weights generated by our gating module. Then, using 'α' indicates whether the input has a prior examination; it is set to 1 if a prior exists, and 0 otherwise. So the fused process is as follows:

$$\hat{F}_{cur} = F_{cur} + \alpha(F_{dy} \odot F_{da}), \tag{20}$$

where $\hat{F}_{cur}$ means the final visual feature after the dynamic fusing process. We can see that F_{dy} can adjust the prior features when the prior examination is available.

IV. EXPERIMENTS

A. Datasets and Metrics

MIMIC-CXR. The MIMIC-CXR dataset [23] consists of 473, 057 images and 227, 835 reports from 63, 478 patients. We utilize the official data split (270,790 for training, 2,130 for validation, and 3,858 for testing). To construct the longitudinal data, we obtain the prior image and report by order different cases according to corresponding timestamps. For each study, we select the most recent historical study as the prior, including consideration of lateral views as well. As a result, the training set contains 146,061 studies with prior exams and 124,729 without; the validation set includes 1,181 with prior exams and 949 without; and the test set consists of 3,153 with prior exams and 705 without.

Longitudinal-MIMIC. This dataset [3] only contains Longitudinal cases. The official data is split as follows: the training set includes 26,156 patients and 92,374 samples, the validation set comprises 203 patients and 737 samples, and the test set contains 266 patients and 2,058 samples.

IU-Xray. The dataset [43] contains a total of 2,955 samples, all are without longitudinal exam. Following the construction strategy of PromptMRG, we reorganized the dataset into 4,168 image and report pairs, all of which are used as the test set.

Evaluation metrics. We employed the same Natural Language Generation (NLG) metrics as those used in previous studies: BLEU scores [44], METEOR [45], ROUGE-L [46]. These metrics were evaluated following the standard evaluation protocol[1]. We also assess clinical efficacy (CE) of reports utilizing the CheXbert [32]. CheXbert can generate disease labels and attributes for each report. Following previous work [8], we map 'positive' and 'uncertain' to the positive label and map 'negative' and 'blank' to the negative label. Then, based on the report classification results, we can calculate the precision (P), recall (R), and F1-score (F1).

Experiment Settings. Our code, based on the Pytorch framework [47], was trained using a single NVIDIA RTX A40 GPU. For all methods used in our comparison, we strictly followed the default configurations reported in their respective original papers without any modifications. To ensure fair and consistent evaluation, we adopted the same training settings as the baseline PromptMRG across all our experiments, including those conducted on different datasets and with various visual backbones. Specifically, all input images were resized to 256 × 256, followed by data augmentations such as 224 × 224 random cropping and random rotation within ±5 degrees. The maximum input text length was set to 100 words. We employed the AdamW optimizer [48] with a weight decay of 0.05. The initial learning rate was set to 5e-5, and all models were trained for 10 epochs with a batch size of 16. Overall loss function combines the Language Modeling (LM) loss $\mathcal{L}_{LM}$ [49] for generation and standard Cross-Entropy (CE) loss $\mathcal{L}_{CE}$ for current and prior disease classification. The total loss $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \mathcal{L}_{LM} + w(\mathcal{L}_{CE}^{cur} + \mathcal{L}_{CE}^{prior}), \tag{21}$$

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

TABLE II

PERFORMANCE COMPARISON ON THREE DIFFERENT VISUAL BACKBONES: RESNET101, SWINT-B, AND EFFICIENTNETV2. VANILLA REPRESENTS THAT VISUAL FEATURES WERE EXTRACTED USING ORIGINAL PRETRAINED MODELS, COUPLED WITH THE FROZEN PRETRAINED BERT FOR TEXT FEATURE EXTRACTION, WITH SUBSEQUENT CONCATENATION OF THESE FEATURES.

Datasets	Visual Backbone	Methods	Clinical Efficacy			Natural Language Generation		
			Precision	Recall	F1	BLEU_4	METEOR	ROUGE_L
MIMIC -CXR	ResNet101	Vanilla	0.395	0.426	0.389	0.108	0.156	0.267
		Ours	**0.411**	**0.449**	**0.421 (+0.032)**	**0.109**	**0.158**	**0.269**
	Efficientv2	Vanilla	0.384	0.430	0.396	0.108	0.157	0.267
		Ours	**0.440**	**0.451**	**0.433 (+0.037)**	**0.111**	**0.160**	**0.270**
	SwinT-B	Vanilla	0.400	0.417	0.392	0.109	0.157	0.269
		Ours	**0.438**	**0.465**	**0.441 (+0.049)**	**0.113**	**0.162**	**0.272**
Longitudinal -MIMIC	ResNet101	Vanilla	0.487	0.544	0.486	0.101	0.151	0.261
		Ours	**0.519**	**0.577**	**0.516 (+0.030)**	**0.105**	**0.155**	**0.266**
	Efficientv2	Vanilla	0.484	0.552	0.487	0.101	0.151	0.264
		Ours	**0.491**	**0.568**	**0.503 (+0.016)**	**0.106**	**0.154**	**0.268**
	SwinT-B	Vanilla	0.506	0.563	0.504	0.102	0.155	0.263
		Ours	**0.539**	**0.584**	**0.527 (+0.023)**	**0.105**	**0.156**	**0.266**

TABLE III

ABLATION STUDY. W/O MSF INDICATES WITHOUT MULTI-STAGE FUSION. W/O DAM INDICATES WITHOUT DIFFERENCE-AWARE MODULE. W/O FAM INDICATES WITHOUT FEATURE ALIGNMENT MODULE. W/O DF INDICATES WITHOUT DYNAMIC FUSION OF BOTH DFAM AND DDAM.

SwinT-B	Model	Clinical Efficacy			Natural Language Generation			
		Precision	Recall	F1	BLEU-1	BLEU-4	METEOR	ROUGE-L
MIMIC -CXR	w/o MSF	0.415	0.419	0.408	0.395	0.107	0.154	0.267
	w/o DAM	0.413	0.417	0.402	0.381	0.100	0.150	0.258
	w/o FAM	0.409	0.455	0.416	0.400	0.111	0.157	0.270
	w/o DF	0.402	0.465	0.421	0.396	0.109	0.158	0.269
	DDaTR	**0.438**	**0.465**	**0.441**	**0.401**	**0.113**	**0.162**	**0.272**
Longitudinal -MIMIC	w/o MSF	0.535	0.556	0.515	0.390	0.103	0.155	0.265
	w/o DAM	0.527	0.556	0.512	0.388	0.101	0.153	0.262
	w/o FAM	0.531	0.567	0.518	0.386	0.103	0.156	0.265
	w/o DF	0.522	0.583	0.521	0.389	**0.106**	0.156	0.266
	DDaTR	**0.539**	**0.584**	**0.527**	**0.396**	0.105	**0.156**	**0.266**

where w is set to 4, consistent with the setting used in PromptMRG.

B. Experimental Results

Comparison on both RRG (MIMIC-CXR, IU-XRay) and LRRG (Longitudinal-MIMIC) benchmarks. We compare our method with a wide range of classical and SOTA methods on MIMIC-CXR, IU-XRay and Longitudinal-MIMIC, and report their results in Tab. I. For the MIMIC-CXR dataset, we follow the evaluation established by RECAP, calculating the Macro score for the CE metrics. For the IU-Xray dataset, we use the same setting as PromptMRG that did not conduct any training or fine-tuning on it, but directly tested our model which trained on MIMIC-CXR. For the Longitudinal-MIMIC dataset, our approach follows the evaluation methodology used in LR2Gen, where the Micro score is calculated for the CE metrics. By comparison, our method achieves SOTA on CE metrics, significantly outperforming other approaches on both RRG and LRRG. We performed a statistical significance analysis on the MIMIC-CXR test set by calculating the p-value of the F1 scores across all longitudinal cases. The result is P_{F1}=0.0415 <0.05 comparing with previous SOTA promptmrg, indicating that the performance improvement of our method is statistically significant. This not only demonstrates that utilizing longitudinal information for report generation tasks yields reports that more closely align with the groundtruth, but also highlights our method's capability to more accurately capture the diagnostic intent of the radiologists. For example, our method with SwinT-B on MIMIC-CXR achieves 0.441 of F1_macro CE score with an absolute improvement of 0.048. Our method with SwinT-B on Longitudinal-MIMIC achieves 0.527 of F1_micro CE score with an absolute improvement of 0.035.

TABLE IV

PERFORMANCE METRICS ON MIMIC-CXR DATASET. COMPARISON WITH LARGE LANGUAGE MODELS.

Model	BLEU-1↑	BLEU-4↑	ROUGE-L↑	METEOR↑	F1-RadGraph↑	RadCliQ↓
RadFM [50]	0.221	0.056	0.205	0.204	0.182	1.687
LLaVA-Med [51]	0.193	0.010	0.139	0.132	0.063	2.155
Med-Flamingo [52]	0.224	0.019	0.145	0.140	0.071	2.116
MedDr [53]	0.322	0.072	0.226	**0.238**	**0.224**	1.538
DDaTR(Ours)	**0.401**	**0.113**	**0.272**	0.162	0.198	**1.141**

TABLE V

PERFORMANCE ON REXRANK DATASET. COMPARISON WITH EXISTING SPECIALIST RRG MODELS.

Model	Year	BLEU-2↑	F1-RadGraph↑	BertScore↑	SembScore↑	1/RadCliQ-V1↑	GREEN↑
RadFM [50]	2023	0.087	0.109	0.313	0.259	0.650	0.185
VLCI [54]	2023	0.136	0.140	0.304	0.305	0.680	0.256
CVT2Dis. [37]	2023	0.126	0.149	0.331	0.329	0.719	0.268
RaDialog [55]	2023	0.127	0.172	0.363	0.387	0.799	0.273
LLM-CXR [56]	2024	0.037	0.046	0.181	0.156	0.516	0.043
MAIRA-2 [57]	2024	0.088	0.131	0.308	0.339	0.694	0.224
CheXagent [58]	2024	0.113	0.148	0.346	0.347	0.741	0.257
MedVersa [59]	2024	0.209	**0.273**	**0.448**	**0.466**	**1.103**	**0.374**
DDaTR(Ours)	2024	**0.210**	0.198	0.362	0.465	0.874	0.293

In the task of RRG, CE metrics are considered more important than NLG metrics. This is because different radiologists exhibit variations in language expression, terminology, and writing styles when composing reports. NLG scores mainly measure the similarity between texts at the sentence level, and minor improvements in NLG scores do not necessarily reflect a more accurate diagnosed of report. However, the precision in describing key observations in report is crucial. CE metrics specifically measure the diagnostic accuracy of reports concerning corresponding attributes. This has been overlooked in some studies which focused on improving NLG scores instead of CE metrics. Therefore, although the NLG scores of our method on the MIMIC-CXR dataset achieved performance just comparable to SOTA methods without showing a significant improvement, excelling in CE metrics could demonstrate our model's superiority and practical utility.

We analyze this as follows: a review of existing methods reveals that higher NLG scores often result from enhancements in the decoder. Thus, under similar data processing and training strategies, the NLG scores largely depend on the ability of the model's decoder, which includes aspects such as the architecture, the tokenizer, and the pre-trained models. Our research primarily focuses on how to encode longitudinal information effectively, aligning input features more closely with medical intent. Consequently, we did not make specific structural modifications to the decoder but instead followed the decoder used in PromptMRG. Furthermore, our zero-shot results on IU X-Ray also outperform previous SOTA, which does not contain longitudinal information. This highlights the robustness and generalizability of our approach and further demonstrates that our model not only substantially improve longitudinal report generation, but also maintain the advanced performance in single-report scenarios.

Comparison on different visual backbones. Our proposed algorithmic framework is adaptable to all multi-scale visual feature extraction networks. To validate its effectiveness, we conducted comparative experiments using three prominent

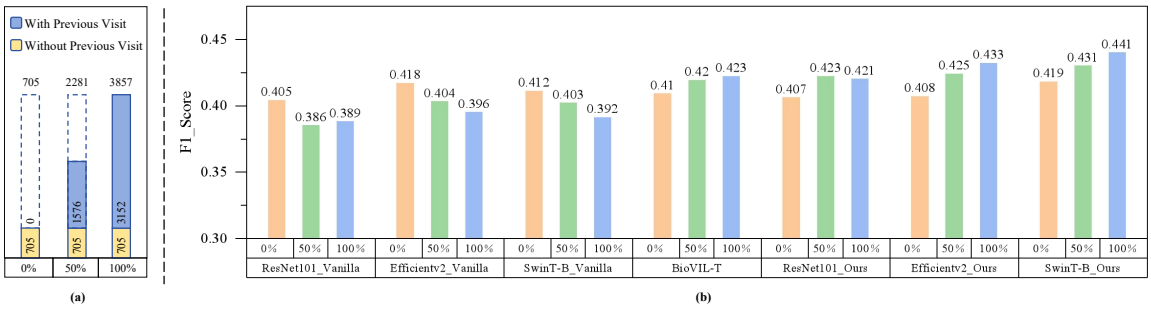


Fig. 7. (a) Illustrates different test settings. Each column represents a specific test setting, with n% indicating the proportion of longitudinal reports used. The total number of samples for each setting is shown at the top of each column. (b) Performance Comparison on different test settings.

visual backbones: ResNet101 [30], SwinT-B [31], and EfficientNetv2 [60]. These networks were selected because they are commonly employed for visual feature extraction in prior RRG research. We compared the performance outcomes between using standard vanilla visual networks and our method under the same baseline framework. In the vanilla approach, visual features were extracted using the original network architectures (e.g., ResNet101) and also fine-tuned during training using the same strategy as ours. For textual features, a pretrained BERT model was also used in a frozen state, and the visual and textual features were subsequently concatenated for downstream tasks. Experimental results, presented in Tab. II, demonstrate that our method outperforms the vanilla encoder across all three backbones by a significant margin.

Comparison on ReXrank. To compare our method with the latest approaches in RRG, we evaluated DDaTR on the official test set of MIMIC-CXR provided by the ReXrank leaderboard [61], which including the comprehensive metrics, such as BERTScore [62], SembScore [32], and GREEN [63]. The results are presented in Tab. V. Our model ranks second on the leaderboard. While MedVersa [59] significantly outperforms all other methods on the leaderboard, it is trained on 10 public datasets and built upon LLM. Due to the substantial differences in both data amount and model size, direct comparison is not entirely fair. Compared to other task-specific LLMs that are explicitly trained on RRG data—such as LLM-CXR [56], CheXagent [58], RaDialog [55] and MAIRA-2 [57]—our method achieves consistently superior performance across all evaluation metrics, further demonstrating the effectiveness of our approach under a more lightweight setting.

Comparison on complexity. To compare model complexity, we report the trainable parameters for ours and PromptMRG as follows: PromptMRG (220M), Ours with Swin-T (360M), Ours with EfficientNetV2 (247M), and Ours with ResNet101 (419M). Notably, when using EfficientNetV2, our model maintains a comparable parameter count (247M) to PromptMRG (220M), yet still achieves significantly better performance as shown in Tab. II. This indicates that the performance gains of our method are not merely attributed to increased model complexity. In fact, in Tab. II Longitudinal-MIMIC, ours with ResNet101 has the highest number of parameters but does not perform as well as the other backbones, further emphasizing that simply increasing model complexity does not guarantee better longitudinal modeling performance. Furthermore, PromptMRG is designed for single-period input, while our method handles longitudinal data, which is acceptable to introduce some additional parameters.

Comparison on different proportions of longitudinal cases. We conducted assessments not only on standard MIMIC-CXR benchmark but also under various settings with different proportions of longitudinal cases. As shown in Fig. 7 (a), we split the testset of MIMIC-CXR. 0% indicates that the test data without longitudinal information. Conversely, 100% means all test cases. At 50%, the test set is composed by randomly selecting half of the reports with prior exams. In the above three testing settings, we evaluated the performance of different encoding methods within the baseline frameworks. Specifically, we tested the effectiveness of three prominent visual backbones using both the standard vanilla approach and our method, alongside another pretrained encoder named BioVIL-T [6], which is designed for LRRG specifically. It is important to note that BioVIL-T utilizes image inputs of 448 x 448 pixels. Our method and other mainstream approaches use a standard input size of 224 x 224 pixels. The larger the image resolution, the richer the image details captured, which naturally provides an additional advantage to BioVIL-T. As illustrated in Fig. 7 (b), despite BioVIL-T's use of larger images, our method still surpasses BioVIL-T. Additionally, as shown in Fig. 7 (b), the vanilla methods underperform in generating multi-period reports. In contrast, our method not only enhances accuracy in single-period report generation but also significantly improves multi-period report generation.

Comparison with medical large language models. Currently, the field of large language models is rapidly evolving, leading to the emergence of numerous substantial medical models aimed at solving a variety of medical image-text tasks. To more comprehensively assess the value of our method, we conducted a comparison with existing medical large language model approaches, including RadFM [50], LLaVA-Med [51], Med-Flamingo [52] and MedDr [53]. The evaluation metrics follow those used in MedDr. As indicated in Tab. IV, in the RRG task, our specialized model demonstrates a distinct advantage over these larger models, particularly in terms of the RadCliQ score [64] (a comprehensive metric that integrates multiple metrics). This suggests that while large models possess robust generalizability and the capability to handle multiple medical image-text tasks, their performance in specific tasks is still inferior to that of dedicated models. Additionally, large models need longer training and inference times, as well as more hardware requirements. In practical applications, particularly for RRG tasks, large models currently cannot fully replace task-specific models.

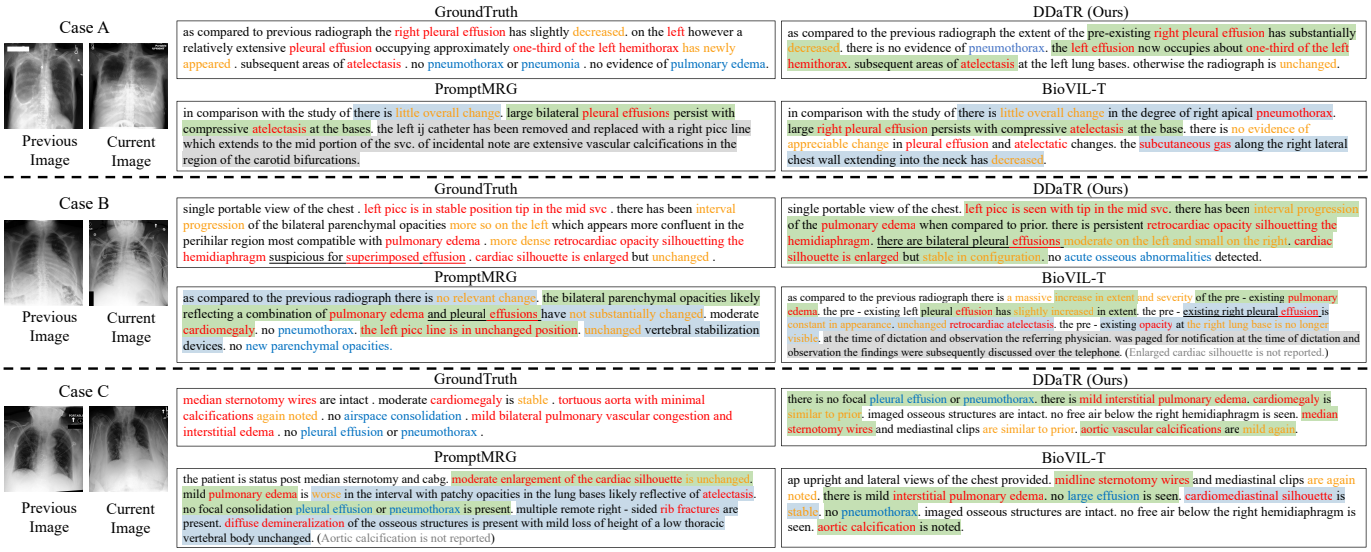


Fig. 8. Case studies comparing radiology report predictions generated by our proposed DDaTR and two baseline methods. **Blue** texts indicates negative observations, the **red** texts indicates positive observations, and the **yellow** texts denotes the progressions. The green background indicates correctly predicted sentences, the blue background denotes incorrectly predicted sentences, and the gray background highlights overstatements.

C. Ablation Study and Case Study

Effect of multi-stage fusion. We initially validated the method utilizing multi-stage fusion. Our analysis compared two approaches: one that aligns multimodal features and detects differences only at the final visual stage, and another that applies feature alignment and difference detection across all visual stages. This comparison helps highlight the difference between our early fusion strategy that integrated within the visual encoder and the late fusion approach adopted by previous LRRG methods. As shown in Tab. III, compared to single-stage feature integration, multi-stage fusion achieved an increase in both the NLG and CE metrics (0.033↑ on F1 of MIMIC-CXR and 0.012↑ on F1 of Longitudinal-MIMIC), highlighting the importance of leveraging multi-level visual features for enhancing longitudinal representation learning.

Effect of DAM and FAM. DAM and FAM represent the configurations where dynamic fusion is removed from the DDAM and DFAM modules, respectively. In our ablation studies, we separately removed DAM and FAM to assess their impacts on the model's performance. As indicated in Tab. III, the FAM led to an F1 increase of 0.025 on the MIMIC-CXR and 0.011 on the Longitudinal-MIMIC. These results demonstrate the effectiveness of FAM in extracting prior visual features that are better aligned with clinical semantics. Similarly, DAM resulted in an an increase of 0.039 in the F1 score on the MIMIC-CXR and 0.018 on the Longitudinal-MIMIC, demonstrating its effectiveness in enhancing the model's ability to capture temporal changes. The comparison between the two experiments indicates that the DAM plays a more critical role, with substantial impact on both the CE and NLG metrics. Moreover, the FAM, which introduces clinical intent of the prior report, shows a notable improvement in CE but has limited effect on NLG. This suggests that the prior report primarily contributes to enhancing diagnostic accuracy.

Effect of dynamic fusion. We evaluated the effect of the dynamic fusion mechanism itself, which adaptively integrates features. As shown in Tab. III, the overall improvement in F1 is primarily attributed to an increase in precision, while the recall remains unchanged. We attribute this to the effectiveness of the dynamic fusion mechanism that enables the model to better handle scenarios where prior information may be missing or noisy, thereby improving diagnostic precision.

Case study. Fig.8 presents a comparison of generated results. As can be observed, our method accurately captures the changes between two period images and provides precise descriptions. Notably, in case A, it can clearly articulate detailed descriptions, such as 'one-third of the left hemithorax'. We analyzed the training data and found that around 10 cases contain this type of description, such as "the effusion occupies approximately one-third of the left hemithorax", "occupying approximately one-third of the left hemithorax". Although the occurrence is relatively rare, our method appears to have captured the associated patterns and is able to generate similar expressions during testing when encountering similar cases. In Case B, we observed that the generated reports still exhibit misalignment issues. Although ours is more accurate than those of the baselines, the ground truth indicates uncertainty regarding effusions, while all models including ours make a definitive statement. This highlights an ongoing challenge in distinguishing between definite and probable diagnoses, which can vary due to differing interpretations of uncertainty among radiologists with different levels of experience. In Case C, we observed that the generated reports still exhibit hallucinations issues. The model occasionally generates additional negative findings or extra information, such as "mediastinal clips". In addition, there are still many cases with diagnostic errors reflecting omissions or false positive issues.

Visualization. To better demonstrate how the model captures and utilizes longitudinal information, we conducted the visualization experiments of our DDATR. Specifically, we generated Grad-CAM [65] visualizations for both the current and previous images according to diagnostic labels, as shown in Fig. 9. We present ten examples, analyzing different diagnostic categories and progressions.

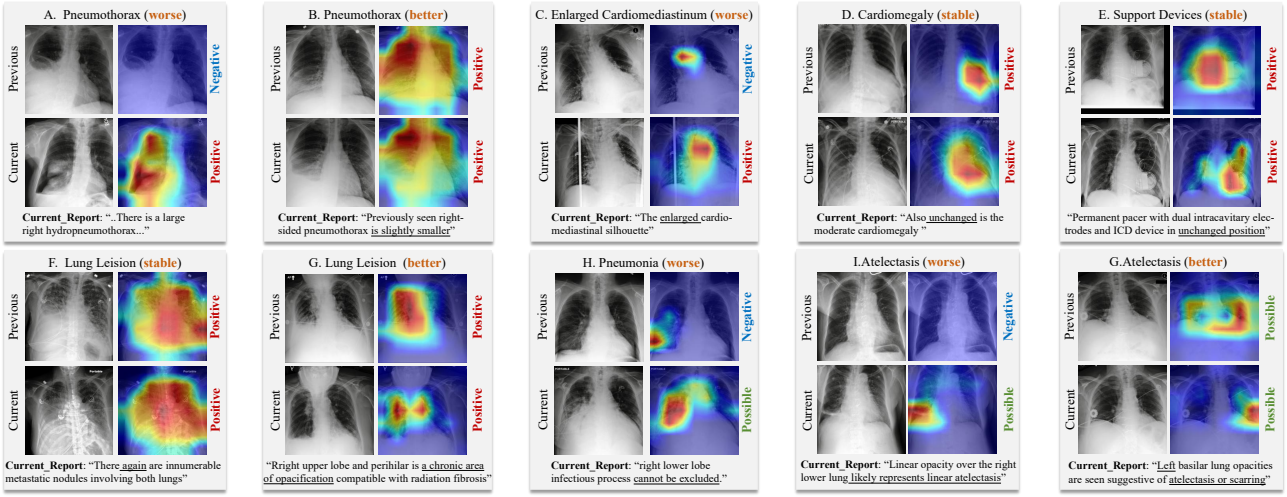


Fig. 9. Grad-CAM visualization examples. Each sub-figure is titled with the case ID, diagnostic class, and progression. The top row displays the previous image and its Grad-CAM map; the bottom row shows the current ones. Diagnostic results (positive, negative, or possible) are provided on the right. At the bottom, the section of current report from the physician are provided for reference, with key information highlighted using '_'.

From these examples, we observe that in Case A and Case I, the model primarily focuses on the current image abnormalities for decision-making, without attending to features from the previous negative exam (less contribution). In contrast, in Case C and Case H, although the previous images also appeared negative for target diseases, the model shows slight attention to the corresponding regions in the previous images when diagnosing current abnormalities. This suggest that the model may tend to reference previous features when making decisions about enlarged (Case C) or when facing uncertainty (Case H). For the remaining examples, where both the previous and current images show positive findings, the model exhibits significant attention to abnormalities across both time points. These observations are consistent with the intended effect of our dynamic feature fusion mechanism.

Also, further analysis of cases where both the previous and current images are positive demonstrate our model's ability to perceive temporal differences between the two longitudinal examinations by using DDAM. For example, in Case B, the visualization shows the slight reduction in right-sided pneumothorax compared to the previous image. In Case C, the results highlights an enlargement of cardiomediastinum.

Moreover, in Case G, the model first focuses on the bilateral lung bases in the previous image, accurately aligning with the input prior report statement: "Patchy bibasilar airspace opacities most likely reflect atelectasis." In the current image, the model's attention shifts to the left lung base, which aligns precisely with the corresponding detail in the current report mentioning "left basilar lung." This strong alignment between generated and reference reports highlights the effectiveness of incorporating prior information (refined, reminded, and guided via DFAM and DDAM) as both informative priors and cross-modal alignment templates, enabling the model to better interpret current findings in context.

V. Conclusion

In this paper, we design a novel dynamic difference-aware temporal residual network (DDaTR) to better capture spatio-temporal correlations among longitudinal information. For temporal perception, our entire network dynamically integrates temporal information based on different prior image features. The residual connections are used to unidirectionally transmit temporal information. For spatial perception, we introduce two modules at each stage of the visual encoder. DFAM employs the entire prior report's features to guide the prior image encoding at each stage. The DDAM leverages differences between two image features to adaptively enhance the source image feature. To better evaluate the comprehensive performance of our network on both single-period and multi-period report generation tasks, we conducted assessments not only on the standard MIMIC-CXR benchmark but also under various settings with different proportions of longitudinal reports. This new evaluation demonstrates the model's versatility in both single-period and multi-period radiology report generation tasks. Extensive experiments and analyses indicate that our method aligns more closely with clinical applications and achieves SOTA results in both RRG and LRRG tasks.

However, our model still has certain limitations that require further investigation and improvement. First, the diagnostic accuracy of current report generation models including ours still does not reach the level of a real radiologist. Issues such as omissions, hallucinations, and misalignments still persist in the generated reports. Therefore, in clinical practice, such models are primarily used to assist radiologists in improving efficiency, with the generated reports still requiring manual review and correction. Secondly, although our method has comprehensive experiments on the X-ray report generation task, its effectiveness on other imaging modalities, such as ultrasound or CT, remains to be further explored and analyzed. Third, each component of our longitudinal framework still holds potential for further optimization, which could lead to further performance gains in future work.

References

[1] A. Esteva, A. Robicquet, B. Ramsundar *et al.*, "A guide to deep learning in healthcare," *Nature medicine*, vol. 25, no. 1, pp. 24–29, 2019.
[2] P. Sloan *et al.*, "Automated radiology report generation: A review of recent advances," *IEEE Reviews in Biomedical Engineering*, 2024.
[3] Q. Zhu, T. S. Mathai *et al.*, "Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports," in *MICCAI*, 2023, pp. 189–198.

[4] F. Dalla Serra, C. Wang, F. Deligianni *et al.*, "Controllable chest x-ray report generation from longitudinal representations," in *EMNLP*, 2023.

[5] F. Wang, S. Du, and L. Yu, "Hergen: Elevating radiology report generation with longitudinal data," *ECCV*, 2024.

[6] S. Bannur *et al.*, "Learning to exploit temporal structure for biomedical vision-language processing," in *CVPR*, 2023, pp. 15 016–15 027.

[7] A. Nicolson, J. Dowling, D. Anderson, and B. Koopman, "Longitudinal data and a semantic similarity reward for chest x-ray report generation," *Informatics in Medicine Unlocked*, vol. 50, p. 101585, 2024.

[8] W. Hou, Y. Cheng, K. Xu, W. Li, and J. Liu, "Recap: Towards precise radiology report generation via dynamic disease progression reasoning," in *EMNLP*, 2023, pp. 2134–2147.

[9] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo *et al.*, "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models," in *CVPR*, 2015, pp. 2641–2649.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

[11] Z. Chen, Y. Song, T. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *EMNLP*, 2020, pp. 1439–1449.

[12] Z. Chen, Y. Shen, Y. Song, and X. Wan, "Cross-modal memory networks for radiology report generation," in *ACL*, 2021, pp. 5904–5914.

[13] H. Qin and Y. Song, "Reinforced cross-modal alignment for radiology report generation," in *Findings of the ACL*, 2022, pp. 448–458.

[14] Z. Wang, L. Liu, L. Wang, and L. Zhou, "Metransformer: Radiology report generation by transformer with multiple learnable expert tokens," in *CVPR*, 2023, pp. 11 558–11 567.

[15] F. Liu, C. You *et al.*, "Auto-encoding knowledge graph for unsupervised medical report generation," *NeurIPS*, vol. 34, pp. 16 266–16 279, 2021.

[16] 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.

[17] Z. Huang, X. Zhang *et al.*, "Kiut: Knowledge-injected u-transformer for radiology report generation," in *CVPR*, 2023, pp. 19 809–19 818.

[18] B. Li, Mingjie Lin *et al.*, "Dynamic graph enhanced contrastive learning for chest x-ray report generation," in *CVPR*, 2023, pp. 3334–3343.

[19] T. Tanida, P. Müller *et al.*, "Interactive and explainable region-guided radiology report generation," in *CVPR*, 2023, pp. 7433–7442.

[20] W. Hou, K. Xu, Y. Cheng *et al.*, "Organ: Observation-guided radiology report generation via tree reasoning," in *ACL*, 2023, pp. 8108–8122.

[21] H. Jin, H. Che *et al.*, "Promptmrg: Diagnosis-driven prompts for medical report generation," in *AAAI*, vol. 38, no. 3, 2024, pp. 2607–2615.

[22] J. T. Wu, N. N. Agu *et al.*, "Chest imagenome dataset for clinical reasoning," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[23] A. E. Johnson *et al.*, "Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs," *arXiv preprint arXiv:1901.07042*, 2019.

[24] X. Yang, M. Ye *et al.*, "Writing by memorizing: Hierarchical retrieval-based medical report generation," in *ACL*, 2021, pp. 5000–5009.

[25] Y. Xue *et al.*, "Multimodal recurrent model with attention for automated radiology report generation," in *MICCAI*, 2018, pp. 457–466.

[26] X. Xie *et al.*, "Attention-based abnormal-aware fusion network for radiology report generation," in *Database Systems for Advanced Applications: DASFAA 2019 International Workshops*, 2019, pp. 448–452.

[27] Y. Miura, Y. Zhang, E. Tsai, C. Langlotz, and D. Jurafsky, "Improving factual completeness and consistency of image-to-text radiology report generation," in *ACL*, 2021, pp. 5288–5304.

[28] F. Liu *et al.*, "Exploring and distilling posterior and prior knowledge for radiology report generation," in *CVPR*, 2021, pp. 13 753–13 762.

[29] V. Ramesh, N. A. Chi, and P. Rajpurkar, "Improving radiology report generation systems by removing hallucinated references to non-existent priors," in *Machine Learning for Health*. PMLR, 2022, pp. 456–473.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[31] Z. Liu, Y. Lin, Y. Cao *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," in *ICCV*, 2021, pp. 10 012–10 022.

[32] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. Lungren, "Combining automatic labelers and expert annotations for accurate radiology report labeling using bert," in *EMNLP*, 2020, pp. 1500–1519.

[33] J. Devlin, M.-W. Chang *et al.*, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *ACL: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 4171–4186.

[34] Z. Wang, L. Liu *et al.*, "R2gengpt: Radiology report generation with frozen llms," *Meta-Radiology*, vol. 1, no. 3, p. 100033, 2023.

[35] L. Huang, W. Wang, J. Chen, and X. Wei, "Attention on attention for image captioning," in *ICCV*, 2019, pp. 4634–4643.

[36] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6070–6080, 2022.

[37] A. Nicolson, J. Dowling, and B. Koopman, "Improving chest x-ray report generation by leveraging warm starting," *Artificial intelligence in medicine*, vol. 144, p. 102633, 2023.

[38] 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.

[39] Z. Yang, J. Wang *et al.*, "Lavt: Language-aware vision transformer for referring image segmentation," in *CVPR*, 2022, pp. 18 155–18 165.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones *et al.*, "Attention is all you need," in *NeurIPS*, 2017, p. 6000–6010.

[41] X. Xie, Y. Cui, T. Tan, X. Zheng, and Z. Yu, "Fusionmamba: Dynamic feature enhancement for multimodal image fusion with mamba," *Visual Intelligence*, vol. 2, no. 1, p. 37, 2024.

[42] P. Huang, H. Ni, Y. Ni *et al.*, "Learnable descriptive convolutional network for face anti-spoofing." in *BMVC*, vol. 2, no. 6, 2022, p. 7.

[43] D. DemnerFushman *et al.*, "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.

[44] K. Papineni, S. Roukos *et al.*, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.

[45] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

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

[47] A. Paszke, S. Gross, F. Massa *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *NeurIPS*, vol. 32, 2019.

[48] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *ICLR*, 2019.

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

[50] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data," *arXiv preprint arXiv:2308.02463*, 2023.

[51] C. Li, C. Wong *et al.*, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *NeurIPS*, vol. 36, 2024.

[52] M. Moor *et al.*, "Med-flamingo: a multimodal medical few-shot learner," in *Machine Learning for Health (ML4H)*. PMLR, 2023, pp. 353–367.

[53] S. He, Y. Nie, Z. Chen, Z. Cai, H. Wang, S. Yang, and H. Chen, "Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning," *arXiv preprint arXiv:2404.15127*, 2024.

[54] W. Chen, Y. Liu, C. Wang *et al.*, "Cross-modal causal intervention for medical report generation," *arXiv preprint arXiv:2303.09117*, 2023.

[55] C. Pellegrini, E. Özsoy, B. Busam, N. Navab, and M. Keicher, "Radialog: A large vision-language model for radiology report generation and conversational assistance," *arXiv preprint arXiv:2311.18681*, 2023.

[56] S. Lee, W. Kim, J. Chang, and J. Ye, "Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation," in *ICLR*, 2024.

[57] S. Bannur, K. Bouzid, D. C. Castro *et al.*, "Maira-2: Grounded radiology report generation," *arXiv preprint arXiv:2406.04449*, 2024.

[58] Z. Chen, M. Varma, J.-B. Delbrouck, M. Paschali *et al.*, "Chexagent: Towards a foundation model for chest x-ray interpretation," in *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

[59] H.-Y. Zhou and Aothers, "Medversa: A generalist foundation model for medical image interpretation," *arXiv preprint arXiv:2405.07988*, 2024.

[60] M. Tan and Q. Le, "Efficientnetv2: Smaller models and faster training," in *ICML*. PMLR, 2021, pp. 10 096–10 106.

[61] X. Zhang, H.-Y. Zhou, X. Yang *et al.*, "Rexrank: A public leaderboard for ai-powered radiology report generation," in *AAAI Bridge Program on AI for Medicine and Healthcare*. PMLR, 2025, pp. 90–99.

[62] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *ICLR*, 2020.

[63] S. Ostmeier, J. Xu, Z. Chen *et al.*, "Green: Generative radiology report evaluation and error notation," in *EMNLP*, 2024, pp. 374–390.

[64] F. Yu, M. Endo, R. Krishnan *et al.*, "Evaluating progress in automatic chest x-ray radiology report generation," *Patterns*, vol. 4, no. 9, 2023.

[65] R. R. Selvaraju *et al.*, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *ICCV*, 2017, pp. 618–626.