LIBRA: LEVERAGING TEMPORAL IMAGES FOR BIOMEDICAL RADIOLOGY ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Radiology report generation (RRG) is a challenging task, as it requires a thorough understanding of medical images, integration of multiple temporal inputs, and accurate report generation. Effective interpretation of medical images, such as chest X-rays (CXRs), demands sophisticated visual-language reasoning to map visual findings to structured reports. Recent studies have shown that multimodal large language models (MLLMs) can acquire multimodal capabilities by aligning with pre-trained vision encoders. However, current approaches predominantly focus on single-image analysis or utilise rule-based symbolic processing to handle multiple images, thereby overlooking the essential temporal information derived from comparing current images with prior ones. To overcome this critical limitation, we introduce Libra, a temporal-aware MLLM tailored for CXR report generation using temporal images. **Libra** integrates a radiology-specific image encoder with a MLLM and utilises a novel Temporal Alignment Connector to capture and synthesise temporal information of images across different time points with unprecedented precision. Extensive experiments show that Libra achieves new state-of-the-art performance among the same parameter scale MLLMs for RRG tasks on the MIMIC-CXR. Specifically, Libra improves the RadCliQ metric by 12.9% and makes substantial gains across all lexical metrics compared to previous models.

1 INTRODUCTION

Radiology reports are a key component of biomedical radiology analysis, designed to enhance clarity and efficiency in medical communication. These structured reports are typically divided into distinct sections, such as *Findings, Impression, Indication, Technique, Comparison*, and *History*(Ganeshan et al., 2018). Serving as the primary medium for radiologists to convey their findings and conclusions from imaging studies like chest X-rays (CXRs), radiology reports play a crucial role in guiding diagnostic and therapeutic decisions for various diseases(Najjar, 2023). However, generating these reports manually is a complex and time-consuming task. Automating radiology report generation (RRG) presents a valuable opportunity to boost radiologist productivity, improve communication, and alleviate burnout (Zhang et al., 2020b). Despite its potential, RRG is a challenging task due to the intricate nature of medical imaging and the need for precise and accurate documentation.

Although the advancements in paradigm models like LLaVA (Liu et al., 2023) and InstructBLIP (Dai et al., 2023), they still exhibit a significant performance gap when applied to RRG tasks. In particular, their effectiveness diminishes in biomedical contexts due to the substantial differences between biomedical and general image-text pairs (Tu et al., 2023; Saab et al., 2024). Consequently, these models often perform at a surface level, akin to a layperson-level understanding, when dealing with specialised medical imaging tasks. Recent advancements have explored continued pre-training of general-purpose foundation models for medical tasks, but these models still fall short in addressing the complexities of medical image analysis (Li et al., 2023a; Chaves et al., 2024; Park et al., 2024).

Harnessing the full potential of multimodal large language models (MLLMs) for understanding and reasoning over multimodal data in the biomedical domain poses a significant challenge, primarily due to the granularity and specificity required in medical image analysis (Wang et al., 2023a). Preliminary attempts to apply MLLMs to RRG have largely concentrated on generating reports from a single image, often neglecting the crucial temporal relationships between images taken at different time points (Zhang et al., 2024b). The MIMIC-CXR Database (Johnson et al., 2019a) reveals that

67% of patients underwent at least two studies at different time intervals, underscoring the necessity of incorporating temporal context in clinical practice. This highlights a critical gap in existing models, which are not equipped to handle the dynamic progression of medical conditions reflected in sequential imaging studies.

In standard clinical practice, radiologists rely heavily on comparing current imaging results with previous studies to assess temporal changes. However, many models inadvertently introduce spurious references to prior examinations that do not exist during inference, leading to inaccurate reports. Recent advancements[1], such as MedVersa (Zhou et al., 2024) and MAIRA-2 (Bannur et al., 2024), can process multiple images simultaneously. These models insert visual tokens from different images at specific points within the textual input, hinging on the LLM to interpret reference information. Additionally, LLaVA-Rad (Chaves et al., 2024) utilised GPT-4V (OpenAI et al., 2024) for radiology report extraction and augmentation, eliminating hallucinations associated with prior image references in the dataset. Despite these efforts, existing models still lack integrated mechanisms for perceiving temporal information within their architectures, which restricts the model's ability to achieve the temporal awareness required for comprehensive reporting. Meanwhile, existing MLLMs typically utilise visual embeddings from the last or penultimate layer of the image encoder (Chen et al., 2023a; Zhang et al., 2024a), which can only capture the global characteristics of the image. However, RRG tasks require not only identifying findings but also capturing subtle details such as severity, extent, and the progression of findings (Sloan et al., 2024). Such fine-grained details are difficult to represent using embeddings from any single layer of the image encoder. (Jiang et al., 2024). Therefore, for the high-granularity demands of RRG tasks, relying solely on representation from a single layer limits the quality of report generation. To tackle these limitations, we aim to enhance the temporal awareness of MLLMs for RRG by tackling two main challenges:

- It is non-trivial to design effective structures in MLLMs to handle prior study citations across various time points in the RRG task.
- The scarcity of effective connectors in MLLMs capable of handling the high-granularity requirements of downstream tasks.

To overcome these challenges, we propose **Libra** (**L**everaging Temporal **I**mages for **B**iomedical **R**adiology **A**nalysis), a novel framework designed to incorporate temporal change information into the RRG task. Libra employs a pre-trained visual transformer, RAD-DINO (Pérez-García et al., 2024), as the image encoder to generate robust image features, which are then refined using a new adapter specifically designed for the temporal awareness, before being fed into the medical large language model (LLM), Meditron (Chen et al., 2023b). Through a two-stage training strategy, Libra demonstrates the potential of MLLMs for specialised radiology applications. Our modular approach integrates state-of-the-art open-source pre-trained models for image and text while using a temporal-aware adapter to align these modalities with the text embedding space. Extensive experiments show that Libra substantially outperforms existing MLLMs in RRG tasks, demonstrating superior cross-modal understanding and reasoning capabilities. Specifically, our contributions include:

- We present **Libra**, the temporal-aware MLLM capable of capturing temporal changes and overcoming the challenge of handling prior study citations, setting a new state-of-the-art performance in RRG tasks on the MIMIC-CXR dataset among MLLMs of the same parameter scale.
- We designed the **Temporal Alignment Connector (TAC)** with two components: the Layerwise Feature Extractor (LFE), which extracts high-granularity image feature embedding from the encoder, and the Temporal Fusion Module (TFM), which integrates temporal references from prior studies. This enhances Libra's ability to capture and utilise temporal information in the RRG task.

2 Libra: Being Aware of Temporal Changes in Chest X-rays Report Generation

2.1 Model Architecture

Libra model follows the common architecture of MLLMs, such as LLaVA, which includes an image encoder, a text decoder (i.e., an LLM pre-trained based on text-only data) and an adapter module for mapping visual representations into text space. In this work, we utilise a frozen biomedical image encoder, RAD-DINO (Pérez-García et al., 2024), a visual transformer extensively pre-trained on medical scans using the DINOv2 image-only self-supervised learning approach (Oquab et al., 2024).

[1]Detailed related work is discussed in Appx. A.1, and our research objectives are explained in Appx. A.2.

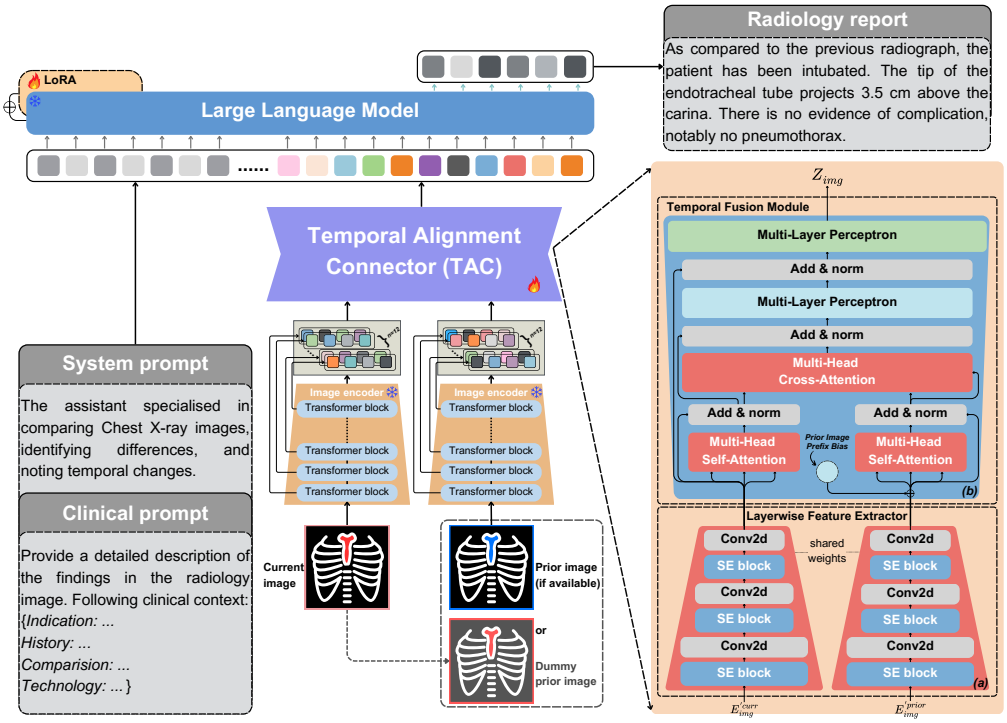


Figure 1: The overall architecture of Libra. The core component of the model is the Temporal Alignment Connector (TAC), designed to process a set of temporal images. It comprises two parts: ***(a)*** the Layerwise Feature Extractor (LFE), and ***(b)*** the Temporal Fusion Module (TFM).

The LLM is deployed by Meditron-7B (Chen et al., 2023b), which builds on Llama-2 (Touvron et al., 2023) and further pre-trained on a specialised medical corpus. However, for the adapter module, we specifically designed a Temporal Alignment Connector (TAC) to capture and integrate temporal information across two images from different time points, as shown in Figure 1. The TAC bridges the image encoder and LLM. It consists of two components: the Layerwise Feature Extractor (LFE), which extracts high-granularity image representations, and the Temporal Fusion Module (TFM), which integrates temporal references from prior studies. This design enables Libra to effectively manage temporal data and enhances its ability to generate accurate and coherent radiology reports.

2.2 Temporal Alignment Connector (TAC): New Adapter Design

To address the challenges of integrating temporal information and aligning high-granularity image features in RRG tasks, we developed the TAC module. It bridges the gap between visual features extracted from multiple temporal snapshots of a patient's CXRs and the language model. The TAC consists of two sub-modules: the LFE, which is to extract the layerwise representation of images, and the TFM, which is to capture the temporal-aware representation between two images.

2.2.1 Layerwise Feature Extractor (LFE)

To leverage abundant feature information in a pre-trained image encoder, we extract all image patch token features from each hidden layer for a given input image. Specifically, RAD-DINO has 12 hidden layers and processes 518 × 518 images into 14 × 14 patches, generating 1,369 patch token sequences per hidden layer for each image. We do not use the $[CLS]$ token. Following this, we obtain patch embeddings of the same dimension from each layer, denoted as $E^{\text{img}} \in \mathbb{R}^{N \times D_{\text{img}}}$, where N is the number of patch tokens and D_{img} is the embedding dimension of the image encoder. Then, we combine them as $E'_{\text{img}} = \{E_i^{\text{img}}\}_{i=1}^{n}$, where n is the number of hidden layers. Drawing from VGG (Simonyan & Zisserman, 2015), we employ a progressive compression strategy to reduce the layer dimension, ensuring that the LFE captures the most relevant features. Initially, we utilise Squeeze-and-Excitation (SE) Networks (Hu et al., 2019), which construct informative features by integrating both spatial and channel-wise information within local receptive fields at each layer.

The SE block is applied to obtain calibrated feature representations, using GELU (Hendrycks & Gimpel, 2023) as the activation function. Next, we employ a specialised pointwise convolution module to align the feature spaces across different layers, using a depthwise 2D convolution with filters and stride of 1, without bias. The dimensions are gradually reduced to better align with the text in the LLM. The compressed features are represented as $A_{\text{img}} \sim Conv2d_j^k(SE_j^k(E'_{\text{img}}))$, where k is the original layer number and j is the layer number after compression. Following the size-reduction pattern of convolutional layers in VGG, the image representations are compressed according to $\{k, j\} \in \{12, 6, 3, 1\}$[2]. Through three stages of progressive compression, we obtain the final representation of image patch features. We define the function $LFE(\cdot)$ to project image features into the same dimension:

$$A_{\text{img}} = LFE(E'_{\text{img}}) \iff A_{\text{img}} = Conv2d_1^3(SE_1^3(A'')) \in \mathbb{R}^{1\times N\times D_{\text{img}}}, \tag{1}$$

where $A'' = Conv2d_3^6(SE_3^6(A'))$, $A' = Conv2d_6^{12}(SE_6^{12}(E'_{\text{img}}))$. The LFE independently processes each image to produce a unified representation for temporal alignment, as ***(a)*** in Figure 1.

2.2.2 Temporal Fusion Module (TFM)

The fusion module is inspired by the transformer decoder. It utilises prior images as auxiliary information to weight the current image, generating a refined, temporal-aware representation suitable for the LLM's latent space. The TFM uses an attention mechanism to dynamically learn relationships between image pairs, enabling the model to adapt to various temporal changes.

Prior Image Prefix Bias: The dataset contains samples with and without a prior image. The patch token features of the current and previous images processed by the LFE are denoted as $A_{\text{img}}^{\text{curr}}$ and $A_{\text{img}}^{\text{prior}}$. If no true prior image exists, we set $A_{\text{img}}^{\text{prior}} = A_{\text{img}}^{\text{curr}}$, effectively using the current image as a dummy prior. To differentiate this case, we introduce a trainable bias b_{prior}. Following the attention mechanism's scaling techniques for adjusting hidden space degrees of freedom with a chi-square distribution (Vaswani, 2017; Wang et al., 2018a), we apply an exponent of $\sqrt[4]{d}$ to the cosine similarity, where d is the hidden dimension of the LLM, to obtain a weighted bias addition to $A_{\text{img}}^{\text{prior}}$:

$$A'^{\text{prior}}_{\text{img}} = A_{\text{img}}^{\text{prior}} + b_{\text{prior}} \cdot \left(\frac{Cos(A_{\text{img}}^{\text{curr}}, A_{\text{img}}^{\text{prior}}) + 1}{2} \right)^{\sqrt[4]{d}}, \tag{2}$$

This nonlinear scaling amplifies higher similarity values, effectively modulating the influence of prior image features. When no true prior image is available, the high similarity score ensures that the effect of the dummy prior is adequately represented. This adjustment prevents samples with a dummy prior image from undergoing redundant rounds of parallel multi-head self-attention during subsequent propagation through the transformer blocks, as shown in Figure 1.

Transformer Block: Our transformer block includes a multi-head cross-attention sub-layer, multi-head self-attention sub-layers, and two multi-layer perceptron (MLP) sub-layers. As shown in Figure 1 ***(b)***. The paired $(A_{\text{img}}^{\text{curr}}, A'^{\text{prior}}_{\text{img}})$ are processed with layer normalization and residual connections:

$$T_{\text{curr}}^{\text{self}} = LayerNorm(A_{\text{img}}^{\text{curr}} + SelfAttn(A_{\text{img}}^{\text{curr}}; A_{\text{img}}^{\text{curr}})), \tag{3}$$

$$T_{\text{prior}}^{\text{self}} = LayerNorm(A'^{\text{prior}}_{\text{img}} + SelfAttn(A'^{\text{prior}}_{\text{img}}; A'^{\text{prior}}_{\text{img}})), \tag{4}$$

$$T_{\text{img}}^{\text{cros}} = LayerNorm(T_{\text{curr}}^{\text{self}} + CorssAttn(T_{\text{curr}}^{\text{self}}; T_{\text{prior}}^{\text{self}})), \tag{5}$$

$$T_{\text{img}}^{\text{out}} = LayerNorm(A_{\text{img}}^{\text{curr}} + MLP_{\text{attn}}(T_{\text{img}}^{\text{cros}})), \tag{6}$$

where MLP_{attn} is a simple neural network composed of two fully connected layers with GELU as the activation function. After that, the features are processed through MLP_{final}, a straightforward neural network consisting of four fully connected layers with the same activation function, but with hidden dimensions matching those of the LLM. We define our fusion module function as $TFM(\cdot)$.

$$Z_{\text{img}} = TFM(A_{\text{img}}^{\text{curr}}, A_{\text{img}}^{\text{prior}}) \iff Z_{\text{img}} = MLP_{\text{final}}(T_{\text{img}}^{\text{out}}), \tag{7}$$

Subsequently, $Z_{\text{img}} \in \mathbb{R}^{N\times d}$ becomes the input sequence that LLM can comprehend, where N is the number of patch tokens and d is the hidden dimension of LLM. The final output is a refined feature representation that encapsulates the temporal evolution of the patient's condition, making it suitable for the language model to generate accurate and contextually aware radiology reports.

[2]Since RAD-DINO has 12 hidden layers, the prime factorisation chain provides the factors as $\{12, 6, 3, 1\}$.

2.3 Prompt Design

To enhance Libra's ability to perceive temporal changes and integrate medical information in the RRG task, we designed a comprehensive prompting strategy comprising a system prompt and detailed clinical prompts, as shown in Figure 1. The system prompt is crafted to enable the LLM to recognise temporal variations: "The assistant specialised in comparing Chest X-ray images, identifying differences, and noting temporal changes." Following this, additional sections of the report, such as {*Indication*}, {*History*}, {*Comparison*}, and {*Technique*}, are incorporated into the clinical instructions (e.g., in Appx. B.4). The clinical prompt to Libra is: "Provide a detailed description of the findings in the radiology image. Following clinical context: {...}." If no clinical instructions are available, the prompt defaults to: "Provide a detailed description of the findings in the radiology image." After tokenizing and embedding the prompt and the answer, we insert the image patch tokens at the specified location, typically between the system prompt and the clinical prompts.

2.4 Temporal-aware Training

In this study, we focus on frontal-view images, either posterior-anterior (PA) or anterior-posterior (AP), and the *Findings* sections of radiology reports, as they contain the most direct clinical observations. We adopt a two-stage training strategy inspired by recent advancements in MLLM fine-tuning techniques (McKinzie et al., 2024). This proposed approach aims to progressively teach the model essential skills, including visual feature alignment and temporal information extraction. Libra's training process involves two stages: temporal feature alignment and downstream task fine-tuning.

In the first stage, the visual encoder and LLM weights are frozen, and the TAC is trained with *Findings* and *Impression* generation, along with CXR-related VQA tasks to extract high-quality image representations and capture temporal changes. In the second stage, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2021) to fine-tune the pre-trained LLM on the *Findings* section generation task, keeping the visual encoder and connector weights frozen. Unlike traditional full fine-tuning, LoRA achieves comparable performance with significantly lower training costs. The detailed training configuration, including learning rate schedules and model parameters, is provided in Appendix B.1.

3 Experiments

3.1 Task and Dataset

Task Description We focus on generating the *Findings* section of radiology reports for a set of frontal CXRs to ensure fair comparison with previous work. The *Findings* section contains descriptions by radiologists of both normal and abnormal findings. Typically, these CXRs are accompanied by additional sections such as *Indication* and *Technique*, which provide the rationale for the study, including clinical history or specific requests from the referring physician. While not directly related to diagnostic interpretation, these sections serve as routine records and assist the model in better understanding temporal changes between images across different periods. Therefore, we incorporate clinical instructions about the current image as prompts to guide Libra to complete the RRG task.

The most common CXR is frontal views, either PA or AP. While other views, such as lateral images, are occasionally used for diagnostic purposes, they mainly serve as supplementary tools to aid in interpreting anatomical structures more comprehensively (Islam et al., 2023). Therefore, the frontal view remains the standard perspective for clinical interpretation of CXRs, and our study is consistent with prior work in RRG tasks, including studies by Chaves et al. (2024) and Hyland et al. (2024). For both the current and prior images, we exclusively use a single frontal view.

Dataset Description To train Libra, we utilise the MIMIC-CXR dataset (Johnson et al., 2019b) and its derivative datasets, Medical-Diff-VQA (Hu et al., 2023) and MIMIC-Ext-*MIMIC-CXR-VQA* (Bae et al., 2023), as shown in the Table 1. All datasets are split according to the official labels to ensure no data leakage. Detailed descriptions of datasets can be found in the Appendix B.2.

Inspired by the multi-stage training techniques commonly employed in recent MLLMs fine-tuning methods (Zhu et al., 2023), we adopt a two-stage training strategy to infuse multimodal capabilities into the model. In the first stage, we leverage a large-scale dataset to train the adapter, primarily focusing on learning the distribution of CXR token embeddings and the general relationships between

Table 1: Datasets used for training and evaluating Libra include statistics on the proportion of samples that contain prior images. In the first stage of training, we utilise the entire dataset, while the second stage focuses on fine-tuning the model on downstream tasks.

Data source	Task Type	# Samples Train (%)	Valid (%)	Test (%)	% Has Prior Train	Valid	Test
MIMIC-CXR	Findings	162 955 (13.43%)	1286 (0.88%)	2461 (2.78%)	58.43	60.11	86.03
	Impression	199 548 (16.45%)	1671 (1.14%)	2343 (2.64%)	64.85	67.09	85.49
Medical-Diff-VQA	Difference	131 563 (10.85%)	16 372 (11.17%)	16 389 (18.48%)	100	100	100
	Abnormality	116 394 (9.59%)	14 512 (9.90%)	14 515 (16.37%)	100	100	100
	Presence	124 654 (10.28%)	15 549 (10.61%)	15 523 (17.51%)	100	100	100
	View	44 970 (3.71%)	5696 (3.89%)	5599 (6.31%)	100	100	100
	Location	67 187 (5.54%)	8510 (5.81%)	8496 (9.58%)	100	100	100
	Level	53 728 (4.43%)	6722 (4.59%)	6846 (7.72%)	100	100	100
	Type	22 067 (1.82%)	2709 (1.85%)	2702 (3.05%)	100	100	100
MIMIC-Ext-*MIMIC-CXR-VQA*	Presence	109 455 (9.02%)	26 153 (17.84%)	4566 (5.15%)	0	0	0
	Anatomy	37 952 (3.13%)	10 210 (6.96%)	1963 (2.21%)	0	0	0
	Attribute	49 948 (4.12%)	13 111 (8.94%)	2578 (2.91%)	0	0	0
	Abnormality	60 692 (5.00%)	16 109 (10.99%)	3199 (3.61%)	0	0	0
	Size	16 000 (1.32%)	4000 (2.73%)	705 (0.80%)	0	0	0
	Plane	7992 (0.66%)	1992 (1.36%)	386 (0.44%)	0	0	0
	Gender	7992 (0.66%)	1992 (1.36%)	396 (0.45%)	0	0	0
Total	Multi-type	1 213 097 (100%)	146 594 (100%)	88 669 (100%)	64.73	49.09	83.67

CXR and text tokens across the MIMIC dataset. The second stage fine-tuning the model on downstream tasks, where the text decoder (i.e., the LLM) is refined to focus on aligning high-granularity information between CXRs and the *Findings* section of radiology reports.

In addition to *Findings* section generation, the temporal feature alignment stage also uses the *Impression* section and VQA tasks. The *Impression* section summarises diagnoses and proposes further investigations (Babar et al., 2021), aiding the alignment between CXRs and their textual descriptions. We use the same system and clinical prompts as for *Findings*, replacing 'Findings' with 'Impression'. For VQA, system prompts remain the same, while clinical prompts adapt to medical-specific questions, guiding the model to generate corresponding captions. These fine-grained VQA tasks expand the MLLM's biomedical vocabulary and improve the alignment of image-text annotations.

3.2 Evaluation Metrics

We evaluate the generated reports using lexical and radiology-specific metrics, following established protocols.. For lexical metrics, we use ROUGE-L (Lin, 2004), BLEU-$\{1, 4\}$ (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005) and BERT (Devlin et al., 2019). For radiology-specific metrics, we use RadGraph-F1 (Jain et al., 2021), RG_{ER} (Delbrouck et al., 2022a), F1-CheXpert (Irvin et al., 2019a), CheXbert vector similarity (Yu et al., 2022), as well as RadCliQ version 0 Yu et al. (2022).

Additionally, we introduce a novel temporal entity F1 score (called $F1_{\text{temp}}$) to assess the model's ability to capture temporal information. Clinical metrics typically emphasise the accuracy of medical findings, prioritising the detection of clinically relevant entities. In contrast, the temporal entity score specifically measures the accuracy of entities related to progression over time described in the report. Appendix B.3 provides detailed descriptions of all metrics and practical analysis of $F1_{\text{temp}}$.

Temporal Entity F1 Following the work of Bannur et al. (2023), we set a reward list that includes common radiology-related keywords associated with temporal changes. Temporal entities are then extracted from both the ground truth (E_{gt}) and the generated reports (E_{gr}). Importantly, no stemming or lemmatization is applied during token processing to preserve the precise description of temporal changes. After extraction, we compute precision (P_{temp}) and recall (R_{temp}), which are subsequently used to calculate the temporal entity score, defined as the harmonic mean of precision and recall (Van Rijsbergen, 1974), also known as the F1 score.

$$P_{\text{temp}} = \frac{|E_{\text{gr}} \cap E_{\text{gt}}| + \epsilon}{|E_{\text{gt}}| + \epsilon}; \quad R_{\text{temp}} = \frac{|E_{\text{gr}} \cap E_{\text{gt}}| + \epsilon}{|E_{\text{gr}}| + \epsilon}; \quad F1_{\text{temp}} = (1+\beta^2) \cdot \frac{P_{\text{temp}} R_{\text{temp}}}{\beta^2 \cdot P_{\text{temp}} + R_{\text{temp}}} \tag{8}$$

where ϵ is a small value, set to a default of 1×10^{-10}, to prevent division by zero (it is also added to the numerator for special cases where no temporal entities are present in the ground truth). The coefficient β controls the balance between precision and recall, with a default value of 1.

3.3 MAIN RESULT

Although MIMIC-CXR provides an "official" test split, strict comparisons with prior studies remain challenging due to differences in test set inclusion criteria and pre-processing steps. For instance, Yu et al. (2022) and Jeong et al. (2023) included only one image per study, resulting in a test set of 1,597 samples, while Tanida et al. (2023) followed the Chest ImaGenome split (Wu et al., 2021b). Such variations in test set distributions can significantly impact the reported results (Park et al., 2024). To ensure fairness, we use a widely adopted test set focused on frontal-view CXRs, aligned with previous studies such as MAIRA-1 (Hyland et al., 2024) and LLaVA-Rad (Chaves et al., 2024) [3].

Additionally, the latest concurrent work M4CXR (Park et al., 2024) employs multi-turn chain-of-thought (Wei et al., 2023) prompting to generate reports, which differs from our task setup. Furthermore, we do not compare with the recent work MAIRA-2, as it is a MLLM specifically designed for grounded radiology report generation task, which incorporates lateral views and prior study reports for each subject within the input prompt. Bannur et al. (2024) emphasises a positive transfer between this task and general report generation, which is also beyond the scope of our study [4].

Taking these considerations into account, we compared our model with state-of-the-art models, including LLaVA-Med (Li et al., 2023a), CheXagent (Chen et al., 2024), GPT-4V (OpenAI et al., 2024), Med-PaLM (Tu et al., 2023), LLaVA-Rad and MAIRA-1. The results are shown in Table 2. Since many of these models are not publicly available, we present their evaluation results as reported in the original sources.

Table 2: Findings generation performance on the official MIMIC-CXR test split. ‡ denotes numbers provided by Chaves et al. (2024), and § represents numbers obtained from Hyland et al. (2024) The best performances in **bold**, and the second-best scores are underlined. '↓' indicates that lower is better. The percentage (%) indicates the improvement over the best existing model.

Metric	LLaVA-Med§	CheXagent‡	GPT-4V‡	Med-PaLM	LLaVA-Rad	MAIRA-1	Libra(%)
Lexical:							
ROUGE-L	27.6	21.5	13.2	27.5	30.6	28.9	**36.7** (19.9%)
BLEU-1	35.4	16.9	16.4	32.3	38.1	39.2	**51.3** (30.9%)
BLEU-4	14.9	4.7	17.8	11.5	15.4	14.2	**24.5** (37.6%)
METEOR	35.3	–	–	–	–	33.3	**48.9** (38.5%)
Clinical:							
RadGraph-F1	19.1	–	–	26.7	–	24.3	**32.9** (23.2%)
RG_{ER}	23.8	20.5	13.2	–	29.4	29.6	**37.6** (27.0%)
$RadCliQ_0$(↓)	3.3	–	–	–	–	3.1	**2.7** (12.9%)
CheXbert vector	36.9	–	–	–	–	44.0	**46.9** (6.59%)
CheXpert-F1:							
Micro-F1-14	42.7	39.3	35.5	53.6	**57.3**	55.7	55.9 (-2.4%)
Macro-F1-14	26.9	24.7	20.4	39.8	39.5	38.6	**40.4** (1.5%)
Micro-F1-5	43.9	41.2	25.8	57.9	57.4	56.0	**60.1** (3.8%)
Macro-F1-5	36.3	34.5	19.6	51.6	47.7	47.7	**53.8** (4.3%)

From Table 2 we can see that, for all traditional lexical metrics, Libra substantially outperforms the baseline models, achieving scores of 36.7 for ROUGE-L, 51.3 and 24.5 for BLEU-1 and BLEU-4 respectively, and 48.9 for METEOR. In clinical metrics, Libra also showed outstanding performance, substantially improving in RadGraph-F1 and its variant RG_{ER}, with scores of 32.9 and 37.6, respectively. Notably, Libra achieved the highest score (i.e. 2.7) on the radiologist-aligned RadCliQ metric. For CheXbert-derived metrics, it also secured the top position with a vector embedding matrix score of 46.9. In the CheXpert 5-class subset, Libra excelled with Micro-F1 and Macro-F1 scores of 60.1 and 53.8, respectively. In the CheXpert 14-class subset, Libra achieved the highest Macro-F1 score of 40.4, while its Micro-F1 score of 55.9 was only slightly below the leading model.

In summary, Libra demonstrated outstanding performance across both sets of metrics, with only minor gaps in one specific measure. The remarkable capabilities of Libra can be attributed to its innovative Temporal Alignment Connector which effectively captures temporal changes in the dataset, enabling the generation of highly accurate and clinically relevant reports.

[3]The test set consists of 2,461 frontal-view image-report samples.

[4]For comparisons with the latest concurrent and non-LLM-based models can be found in Appendix D.

4 Ablation Study

To thoroughly assess the impact of various optimisation components, we conducted a series of ablation experiments, evaluating different modules and dataset expansions within Libra. All ablation models were tested on the MIMIC-CXR official test split for the *Findings* section generation task, using identical hyperparameter settings during training and inference.

Q1: Does incorporating temporal information bring positive effects to Libra in RRG tasks?

Temporal information is embedded in paired images and referenced in the corresponding radiology reports, reflecting changes over time through references to previous symptoms and descriptions of their progression. As indicated in Table 1, 86% of the test data includes true prior images. To assess whether Libra can effectively perceive and utilise temporal information during inference, we conducted separate evaluations using either true prior images or only dummy prior images. When true prior images were unavailable, the model treated the current image as a dummy prior image. As shown in Table 3, Libra demonstrated substantial enhancements across all metrics when true prior images were used as references. Specifically, the improvement in clinical scores exceeds that of lexical scores, indicating that temporal information is crucial for generating high-quality medical reports, beyond merely improving linguistic fluency. Notably, the $F1_{temp}$ score shows the most significant impact, with a difference of **7.41%**. This highlights Libra's capability to effectively leverage temporal changes provided by true prior images, thereby enhancing the quality of the generated *Findings* section and overall performance in the RRG task.

Table 3: Ablation results for Libra using true prior images or only dummy prior images. Values in (%) indicate the percentage improvement.

Metric	Libra	
	w/ true	w/ dummy(%)
Lexical:		
ROUGE-L	36.66	36.17 (-1.34%)
BLEU-1	51.25	51.20 (-0.10%)
BLEU-4	24.54	24.33 (-0.86%)
METEOR	48.90	48.69 (-0.43%)
BERTScore	62.50	61.94 (-0.90%)
$F1_{temp}$	35.34	32.72 (-7.41%)
Clinical:		
RadGraph-F1	32.87	32.42 (-1.37%)
RG_{ER}	37.57	36.92 (-1.73%)
$RadCliQ_0$(↓)	2.72	2.76 (-1.47%)
CheXbert vector	46.85	46.31 (-1.15%)
CheXpert-F1:		
Micro-F1-14	55.87	55.25 (-1.11%)
Macro-F1-14	40.38	40.15 (-0.57%)
Micro-F1-5	60.07	58.93 (-1.90%)
Macro-F1-5	53.75	52.61 (-2.12%)

Q2: Does the Temporal Alignment Connector effectively improve model performance?

Table 4: Results of ablation experiments for the Temporal Alignment Connector. '↓' indicates that lower is better. Values in (%) indicate the percentage decrease compared with the Libra-1.

Metric	Libra-1	w/o TFM	w/o LFE	w/o PIPB	w/o TAC
Lexical:					
ROUGE-L	27.56	27.33 (-0.85%)	27.21 (-1.27%)	27.43 (-0.48%)	26.17 (-5.04%)
BLEU-1	34.84	34.17 (-1.92%)	34.21 (-1.82%)	34.60 (-0.67%)	33.03 (-5.20%)
BLEU-4	11.51	11.13 (-3.33%)	11.11 (-3.47%)	11.43 (-0.73%)	10.02 (-12.98%)
METEOR	35.50	35.06 (-1.24%)	34.96 (-1.52%)	35.28 (-0.62%)	33.98 (-4.28%)
BERTScore	55.87	55.60 (-0.49%)	55.49 (-0.69%)	55.74 (-0.23%)	54.63 (-2.22%)
$F1_{temp}$	26.63	25.96 (-2.51%)	26.21 (-1.57%)	26.58 (-0.18%)	25.39 (-4.65%)
Clinical:					
RadGraph-F1	22.52	2.20 (-1.42%)	22.03 (-2.19%)	22.35 (-0.74%)	21.51 (-4.48%)
RG_{ER}	27.32	26.89 (-1.59%)	26.72 (-2.19%)	27.09 (-0.84%)	25.97 (-4.96%)
$RadCliQ_0$ (↓)	3.10	3.12 (-0.65%)	3.12 (-0.65%)	3.11 (-0.32%)	3.15 (-1.61%)
CheXbert vector	42.02	41.57 (-1.07%)	41.37 (-1.54%)	41.92 (-0.11%)	40.93 (-2.59%)
CheXpert-F1:					
Micro-F1-14	52.48	51.74 (-1.42%)	51.68 (-1.53%)	52.13 (-0.67%)	51.13 (-2.57%)
Macro-F1-14	36.87	36.04 (-2.25%)	36.12 (-2.03%)	36.14 (-1.97%)	35.85 (-2.76%)
Micro-F1-5	56.63	55.37 (-2.23%)	55.79 (-1.49%)	55.87 (-1.34%)	54.51 (-3.74%)
Macro-F1-5	49.33	47.76 (-3.18%)	47.82 (-3.06%)	47.98 (-2.75%)	47.22 (-4.28%)

To evaluate the impact of TAC on Libra's performance in the RRG task, we initialised a model with the RAD-DINO (Pérez-García et al., 2024) image encoder, TAC, and Meditron (Chen et al., 2023b) as the LLM. A baseline experiment (Libra-1) was conducted by fine-tuning only the TAC for the *Findings* generation task. As shown in Table 4, we performed ablation studies by progressively removing different TAC components, including TFM, LFE, the Prior Image Prefix Bias (PIPB), and the entire TAC. Removing TFM restricted the model to processing only the current image, using a configuration similar to LLaVA (Liu et al., 2023) but with a four-layer MLP to match the image representation with the LLM's hidden dimensions. Without LFE, the model follows the LLaVA setup,

utilising the penultimate layer of the image encoder. The results indicate that removing any TAC submodule leads to a decline in all metrics compared to Libra-1. Removing TFM caused a notable drop in the $F1_{temp}$ score (↓>2%), emphasising its role in capturing temporal information. Removing LFE especially decreases RadGraph-related scores, highlighting its importance in extracting detailed image features. The removal of PIPB impacts clinical scores more than lexical scores. Finally, removing the entire TAC results in significant declines across all metrics, underscoring its critical role in integrating image details and temporal information[5].

Q3: Are additional *Impression* and VQA datasets needed during the feature alignment?

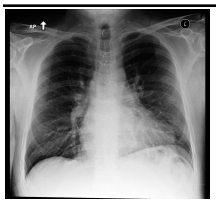
Table 5: Ablation experiments of dataset expansion in Libra. △ indicates better performance at the same stage, while ▼ indicates the opposite.

Metric	Stage: 1		Stage: 2	
	Libra	*Libra-f*	**Libra**	*Libra-f*
Lexical:				
ROUGE-L	27.27	27.56△	36.66	35.31▼
BLEU-1	41.24	34.84▼	51.25	49.92▼
BLEU-4	13.59	11.51▼	24.54	23.05▼
METEOR	39.44	35.50▼	48.90	47.99▼
BERTScore	56.00	55.87▼	62.50	61.28▼
$F1_{temp}$	24.80	26.63△	35.34	33.52▼
Clinical:				
RadGraph-F1	20.45	22.52△	32.87	30.77▼
RG_{ER}	25.19	27.32△	37.27	35.44▼
$RadCliQ_0$ (↓)	3.31	3.10△	2.72	2.83▼
CheXbert vector	35.33	42.02△	46.85	45.32▼
CheXpert-F1:				
Micro-F1-14	43.63	52.48△	55.87	54.11▼
Macro-F1-14	25.68	36.87△	40.38	37.16▼
Micro-F1-5	49.75	56.63△	60.07	58.76▼
Macro-F1-5	40.40	49.33△	53.75	51.99▼

To evaluate the impact of dataset expansion during the first stage on Libra's performance, we trained a model (Libra-f) using only the *Findings* data in the first stage, while Libra incorporated additional *Impression* and VQA data for alignment, in Table 5. After the first stage, Libra showed superior performance in lexical metrics compared to Libra-f, but a slight decline in clinical metrics. This can be attributed to the inclusion of VQA tasks, which encouraged the model to focus on more fine-grained and grounded information. While VQA emphasises the detailed description of the single symptom, the *Findings* section generation task requires a comprehensive overview of the CXR, encompassing multiple normal and abnormal findings. This shift in focus also impacted the $F1_{temp}$ score, as each finding typically involves its own temporal changes. Additionally, the reduction in the number of generated findings entities led to a decrease in the number of identified temporal entities. In the second stage, fine-tuned on the *Findings* dataset, Libra achieved the best results across all metrics, indicating that incorporating additional datasets in the first stage enhances Libra's understanding and reasoning of CXRs.

5 Performance Analysis

We used CXRs from the official test split to generate *Findings* sections and perform a qualitative analysis of Libra. Figure 2 illustrates a case where no true prior image was available for reference. For example, while the ground truth report only mentioned the presence of "sternal wires," Libra not only identified their presence but also provided details on the specific type. This demonstrates Libra can deliver more precise clinical information by not only recognizing conditions but also distinguishing the specific types of surgical interventions.



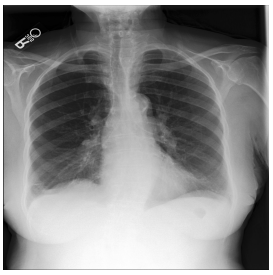
Original report	The lungs are well expanded and clear. The cardiomediastinal silhouette, hilar contours, and pleural surfaces are normal. No pleural effusion or pneumothorax is present. Sternal wires are intact.
Libra	The lungs are well expanded and clear. The cardiomediastinal silhouette, hilar contours, and pleural surfaces are normal. No pleural effusion or pneumothorax is present. **Median sternotomy** wires are intact.

Figure 2: Each identified radiological feature is highlighted. Surgical intervention type in **bold**.

For the samples with true prior images, Libra accurately describes the presence and changes in radiological conditions. As shown in the Figure 3, the current image revealed additional findings of pleural effusion and pneumonia compared to the prior image. The model effectively detected these progressive changes, provided detailed descriptions, and suggested potential causes for further investigation. Notably, it explicitly mentioned the comparative information relative to the prior image, offering a clear understanding of the temporal changes.

[5]For more ablation studies, including the impact of radiology-specific pre-trained models, see Appendix C.

Current image



Prior image

Original report (current image)	In comparison with the study of ___, there is again evidence of a superior mediastinal mass consistent with the known thyroid enlargement that is better seen on the recent CT scan. Continued bibasilar opacification most likely represents atelectasis, though in the appropriate clinical setting, superimposed pneumonia would have to be considered. The cardiac silhouette is within normal limits, though there may be some mild elevation of pulmonary venous pressure. Small pleural effusions are probably present bilaterally.
Libra	As compared to ___ radiograph, cardiomediastinal contours are stable. New patchy and linear opacities have developed at the lung bases, right greater than left, and may reflect patchy atelectasis, aspiration, or developing infectious pneumonia. Small pleural effusions are present bilaterally.
w/ dummy prior image	Cardiac silhouette is upper limits of normal in size and accompanied by pulmonary vascular congestion and minimal interstitial edema. Patchy bibasilar opacities are present, and could reflect patchy atelectasis, aspiration, or developing infectious pneumonia. Followup radiographs may be helpful in this regard. Small bilateral pleural effusions are also demonstrated.
Swap order	Cardiomediastinal contours are normal. Previously present bibasilar opacities have nearly **completely resolved** with minimal residual linear opacification remaining at the right lung base. **No** new areas of consolidation are identified, and there are **no** pleural effusions.
Original report (prior image)	The heart is normal in size. The mediastinal and hilar contours appear stable. The aortic arch is calcified. There is no definite pleural effusion or pneumothorax. Streaky opacity in the right costophrenic angle suggests minor atelectasis, which has decreased. There is similar mild spinal curvature.

Figure 3: Radiological features are highlighted. Descriptions of temporal changes are marked. Key highlights from the swapped order are in **bold**. Heatmap analysis can be found in Appendix E.

To further assess the model's robustness, we conducted a more detailed analysis of this case. First, we generated a report using only the dummy prior image as a reference. While Libra successfully detected abnormal findings, it failed to specify changes relative to the previous image, describing only the current state without indicating disease progression. Moreover, it provided less clarity in distinguishing the distribution and severity of abnormalities. This suggests that, without a true prior image for comparison, the model struggles to accurately describe the progression of findings.

Evaluating Temporal Consistency in Libra's Report Generation To investigate the model's ability to capture temporal changes, we swapped the order of the images, using the true prior image as the current image and vice versa. Interestingly, the generated report indicated an improvement in the patient's condition, with most abnormalities resolved and no new findings detected, which is contrary to the ground truth of the original current image report. Moreover, the report generated from the swapped images closely resembled the original report of the true prior image, as shown at the bottom of Figure 3, which described a stable or improving condition for the patient. This striking observation demonstrates that Libra can effectively utilise both current and prior studies to generate accurate reports for assessing changes, simulating the conditions of standard clinical practice.

6 Conclusion

In this study, we presented Libra, a temporal-aware multimodal large language model specifically designed for chest X-rays report generation tasks. Trained exclusively on the open-access MIMIC-CXR dataset (Johnson et al., 2019b), Libra utilises a two-stage training framework that leverages a radiology-specific pre-trained image encoder and language model, connected through a specially designed Temporal Alignment Connector to bridge visual and textual modalities. By identifying observations from current CXR and referencing prior scan, Libra generates radiology reports with substantial improvements across all metrics compared to existing models of the same parameter scale. Qualitative analysis in the case study highlights Libra's ability to effectively utilise the temporal relationships between current and prior images, addressing the challenge of hallucinations related to prior study references in RRG tasks. Libra sets a new paradigm for MLLMs in multimodal medical AI research. In the future, we aim to further enhance Libra's clinical applicability and accuracy, paving the way for the development of general-purpose AI models in the medical domain. For a detailed discussion of Libra's limitations and future work, please refer to Appendix A.3.

REFERENCES

Zaheer Babar, Twan van Laarhoven, Fabio Massimo Zanzotto, and Elena Marchiori. Evaluating diagnostic content of ai-generated radiology reports of chest x-rays. *Artificial Intelligence in Medicine*, 116:102075, 2021. URL https://doi.org/10.1016/j.artmed.2021.102075.

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric I-Chao Chang, Tackeun Kim, and Edward Choi. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images, 2023. URL https://arxiv.org/abs/2310.18652.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. Learning to exploit temporal structure for biomedical vision-language processing, 2023. URL https://arxiv.org/abs/2301.04558.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Fabian Falck, Ozan Oktay, Anja Thieme, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. Maira-2: Grounded radiology report generation, 2024. URL https://arxiv.org/abs/2406.04449.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66: 101797, 2020.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation, 2024. URL https://arxiv.org/abs/2403.08002.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, January 2023a. ISSN 2731-5398. doi: 10.1007/s11633-022-1369-5. URL http://dx.doi.org/10.1007/s11633-022-1369-5.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. Meditron-70b: Scaling medical pretraining for large language models, 2023b. URL https://arxiv.org/abs/2311.16079.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5904–5914, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.459. URL https://aclanthology.org/2021.acl-long.459.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer, 2022. URL https://arxiv.org/abs/2010.16056.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. Chexagent: Towards a foundation model for chest x-ray interpretation, 2024. URL https://arxiv.org/abs/2401.12208.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL https://arxiv.org/abs/2204.02311.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. URL https://arxiv.org/abs/2305.06500.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.319. URL https://aclanthology.org/2022.findings-emnlp.319.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis P. Langlotz. Improving the factual correctness of radiology report generation with semantic rewards, 2022b. URL https://arxiv.org/abs/2210.12186.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Dhakshinamoorthy Ganeshan, Phuong-Anh Thi Duong, Linda Probyn, Leon Lenchik, Tatum A McArthur, Michele Retrouvey, Emily H Ghobadi, Stephane L Desouches, David Pastel, and Isaac R Francis. Structured reporting in radiology. *Academic radiology*, 25(1):66–73, 2018. URL https://doi.org/10.1016/j.acra.2017.08.005.

Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. Complex organ mask guided radiology report generation, 2023. URL https://arxiv.org/abs/2311.02329.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. URL https://arxiv.org/abs/1606.08415.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. RECAP: Towards precise radiology report generation via dynamic disease progression reasoning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2134–2147, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.140. URL https://aclanthology.org/2023.findings-emnlp.140.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. Organ: Observation-guided radiology report generation via tree reasoning, 2023b. URL https://arxiv.org/abs/2306.06466.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. Squeeze-and-excitation networks, 2019. URL https://arxiv.org/abs/1709.01507.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 4156–4165, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599819. URL https://doi.org/10.1145/3580305.3599819.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. Kiut: Knowledge-injected u-transformer for radiology report generation, 2023. URL https://arxiv.org/abs/2306.11345.

Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. Maira-1: A specialised large multimodal model for radiology report generation, 2024. URL https://arxiv.org/abs/2311.13668.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019a. URL https://arxiv.org/abs/1901.07031.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 590–597, 2019b.

S. K. M Shadekul Islam, MD Abdullah Al Nasim, Ismail Hossain, Md Azim Ullah, Kishor Datta Gupta, and Md Monjur Hossain Bhuiyan. Introduction of medical imaging modalities, 2023. URL https://arxiv.org/abs/2306.01022.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports, 2021. URL https://arxiv.org/abs/2106.14463.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. Multimodal image-text matching improves retrieval-based chest x-ray report generation, 2023. URL https://arxiv.org/abs/2303.17579.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models, 2024. URL https://arxiv.org/abs/2310.08825.

Alistair E W Johnson, David J Stone, Leo A Celi, and Tom J Pollard. The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39, 2018. URL https://doi.org/10.1093/jamia/ocx084.

Alistair E. W. Johnson, Tom J. Pollard, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih ying Deng, Yifan Peng, Zhiyong Lu, Roger G. Mark, Seth J. Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019a. URL https://arxiv.org/abs/1901.07042.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019b. URL https://doi.org/10.13026/4jqj-jw95.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023a. URL https://arxiv.org/abs/2306.00890.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation, 2023b. URL https://arxiv.org/abs/2303.10323.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pp. 249–269. PMLR, 2019a.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In Finale Doshi-Velez, Jim Fackler, Ken Jung, David Kale, Rajesh Ranganath, Byron Wallace, and Jenna Wiens (eds.), *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pp. 249–269. PMLR, 09–10 Aug 2019b. URL https://proceedings.mlr.press/v106/liu19a.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. URL https://arxiv.org/abs/2304.08485.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, Anton Belyi, Haotian Zhang, Karanjeet Singh, Doug Kang, Ankur Jain, Hongyu Hè, Max Schwarzer, Tom Gunter, Xiang Kong, Aonan Zhang, Jianyu Wang, Chong Wang, Nan Du, Tao Lei, Sam Wiseman, Guoli Yin, Mark Lee, Zirui Wang, Ruoming Pang, Peter Grasch, Alexander Toshev, and Yinfei Yang. Mm1: Methods, analysis and insights from multimodal llm pre-training, 2024. URL https://arxiv.org/abs/2403.09611.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5288–5304, Online, June 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.416. URL https://aclanthology.org/2021.naacl-main.416.

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation, 2021b. URL https://arxiv.org/abs/2010.10042.

Reabal Najjar. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*, 13(17):2760, 2023. URL https://www.mdpi.com/2075-4418/13/17/2760.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. Progressive transformer-based generation of radiology reports. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 2824–2832, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-emnlp.241. URL https://aclanthology.org/2021.findings-emnlp.241.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng,

Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. URL https://arxiv.org/abs/2304.07193.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pp. 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Jonggwon Park, Soobum Kim, Byungmu Yoon, Jihun Hyun, and Kyoyun Choi. M4cxr: Exploring multi-task potentials of multi-modal large language models for chest x-ray interpretation, 2024. URL https://arxiv.org/abs/2408.16213.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, Maria Wetscherek, Noel Codella, Stephanie L. Hyland, Javier Alvarez-Valle, and Ozan Oktay. Rad-dino: Exploring scalable medical image encoders beyond text supervision, 2024. URL https://arxiv.org/abs/2401.10815.

Han Qin and Yan Song. Reinforced cross-modal alignment for radiology report generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 448–458, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.38. URL https://aclanthology.org/2022.findings-acl.38.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models, 2020. URL https://arxiv.org/abs/1910.02054.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathon Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. Capabilities of gemini models in medicine, 2024. URL https://arxiv.org/abs/2404.18416.

Santosh Sanjeev, Fadillah Adamsyah Maani, Arsen Abzhanov, Vijay Ram Papineni, Ibrahim Almakky, Bartłomiej W. Papież, and Mohammad Yaqub. Tibix: Leveraging temporal information for bidirectional x-ray and report generation, 2024. URL https://arxiv.org/abs/2403.13343.

Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeffrey Dalton, and Alison Q O'Neil. Controllable chest x-ray report generation from longitudinal representations, 2023. URL https://arxiv.org/abs/2310.05881.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, pp. 1–20, 2024. ISSN 1941-1189. doi: 10.1109/rbme.2024.3408456. URL http://dx.doi.org/10.1109/RBME.2024.3408456.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117.

Xiao Song, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Pengxu Wei. Cross-modal contrastive attention model for medical report generation. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na (eds.), *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 2388–2397, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL https://aclanthology.org/2022.coling-1.210.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7433–7442, June 2023. URL https://doi.org/10.1109/CVPR52729.2023.00718.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, Basil Mustafa, Aakanksha Chowdhery, Yun Liu, Simon Kornblith, David Fleet, Philip Mansfield, Sushant Prakash, Renee Wong, Sunny Virmani, Christopher Semturs, S Sara Mahdavi, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Karan Singhal, Pete Florence, Alan Karthikesalingam, and Vivek Natarajan. Towards generalist biomedical ai, 2023. URL https://arxiv.org/abs/2307.14334.

Cornelis Joost Van Rijsbergen. Foundation of evaluation. *Journal of Documentation*, Volume 30 (Issue 4):365–373, 1974. doi: 10.1108/eb026584. URL https://doi.org/10.1108/eb026584.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. URL https://doi.org/10.48550/arXiv.1706.03762.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie fu. Pre-trained language models in biomedical domain: A systematic survey, 2023a. URL https://arxiv.org/abs/2110.05006.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks, 2018a. URL https://arxiv.org/abs/1711.07971.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays, 2018b. URL https://arxiv.org/abs/1801.04334.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens, 2023b. URL https://arxiv.org/abs/2304.02211.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Joy Wu, Nkechinyere Agu, Ismini Lourentzou, Arjun Sharma, Joseph Paguio, Jasper Seth Yao, Edward Christopher Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset. *Physio Net*, 2021a.

Joy T. Wu, Nkechinyere N. Agu, Ismini Lourentzou, Arjun Sharma, Joseph A. Paguio, Jasper S. Yao, Edward C. Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, Leo A. Celi, and Mehdi Moradi. Chest imagenome dataset for clinical reasoning, 2021b. URL https://arxiv.org/abs/2108.00316.

Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80:102510, August 2022. ISSN 1361-8415. doi: 10.1016/j.media.2022.102510. URL http://dx.doi.org/10.1016/j.media.2022.102510.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*, 2022. doi: 10.1101/2022.08.30.22279318. URL https://www.medrxiv.org/content/early/2022/08/31/2022.08.30.22279318.

Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, and Jian Wu. Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant, 2024. URL https://arxiv.org/abs/2403.04290.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey, 2024a. URL https://arxiv.org/abs/2304.00685.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024b. URL https://arxiv.org/abs/2303.00915.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020a. URL https://arxiv.org/abs/1904.09675.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph, 2020b. URL https://arxiv.org/abs/2002.08277.

Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling, 2023. URL https://arxiv.org/abs/2301.13155.

Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J. Topol, and Pranav Rajpurkar. A generalist learner for multifaceted medical image interpretation, 2024. URL https://arxiv.org/abs/2405.07988.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023. URL https://arxiv.org/abs/2304.10592.

A Appendix A

A.1 Related Work

Radiology Report Generation Several efforts have been made to develop models tailored for radiology report generation (RRG), driven by the need to address the long-tail distribution of observations in chest X-rays and provide fine-grained descriptions of findings. This task has gained prominence as an important objective for the designed system (Wang et al., 2018b).

Such systems inherently require a language generation component. Initially, recurrent neural networks were used (Liu et al., 2019a), but these have recently been replaced by Transformer architectures (Miura et al., 2021b; Chen et al., 2022), including large language models (LLMs) such as PaLM (Chowdhery et al., 2022) and Vicuna-7B (Chiang et al., 2023).

To ensure clinical accuracy in generated text, some studies have moved beyond standard language modelling losses, instead employing reinforcement learning (RL) to optimise for rewards that prioritise "clinically relevant" features, such as the inclusion of specific findings (Liu et al., 2019b; Irvin et al., 2019b) or maintaining logical consistency (Miura et al., 2021a; Delbrouck et al., 2022a). However, these methods often depend on external models like CheXbert (Smit et al., 2020) or RadGraph (Jain et al., 2021) to extract clinical entities, adding complexity to the optimisation process.

With the advancement of LLMs, an increasing number of studies have demonstrated that plain autoregressive language modelling can achieve substantial results. However, we acknowledge that the benefits of more complex training objectives or RL-based approaches may be complementary. Consequently, research on leveraging temporal information in RRG tasks can be broadly categorised into two types: LLM-based methods and non-LLM-based methods.

non-LLM-based Model These architectures are typically transformer encoder-decoder models and their variants. They are often composed of multiple modules that require separate training. Meanwhile, they handle single and double image inputs by symbolically differentiating tasks and employing two distinct neural network architectures: one designed for single-image inputs and another for double-image inputs.

Serra et al. (2023) utilises symbolic alignment in its Longitudinal Projection Module and employs a separately trained BERT-based (Devlin et al., 2019) text generator. RECAP (Hou et al., 2023a) implements a transformer encoder-decoder with symbolic task differentiation and two-stage training, beginning with classification tasks followed by report generation. TiBiX (Sanjeev et al., 2024) adopts a transformer model with causal attention layers and incorporates learnable padding tokens to handle cases without prior images. BioViL-T (Bannur et al., 2023) is a self-supervised vision-language training framework that features a CNN–Transformer hybrid multi-image encoder trained jointly with a BERT-based text model.

On one hand, the difference in model parameter sizes, and on the other, as LLM-based models generally outperform other types of models in the RRG task, papers on non-LLM-based models or those using small language models typically do not compare their methods with LLM-based approaches. At the same time, we conducted comparisons and discussions to reaffirm this point, as detailed in Appendix D.2.

LLM-based Model Current LLM-based models have achieved notable success in the RRG task, largely benefiting from visual instruction tuning (Liu et al., 2023). Structurally, these models (Li et al., 2023a; Chaves et al., 2024; Hyland et al., 2024; Zhou et al., 2024; Park et al., 2024) typically consist of an image encoder and an adapter that bridges the connection to the LLM. These adapters often rely on single-layer hidden representations (e.g., the last or penultimate layer) from pre-trained image encoders, limiting their ability to simultaneously integrate information from multiple images.

In end-to-end training, LLM-based models handle multiple image inputs by concatenating them with textual prompts to form a composite input for the LLM. For example, the input format is "<Current Image Placeholder> + <Prior Imgae Placeholder> + <Prompt>".

Although our model adopts an end-to-end training paradigm, it does not rely on the vanilla approach to handle multiple images. Instead, we designed a novel adapter called the Temporal Alignment

Connector (TAC). On one hand, TAC leverages all hidden layer features from the image encoder, setting it apart from the default configurations used in previous works. On the other hand, TAC models temporal information directly rather than simply concatenating two images in the latent space of the LLM.

Radiological Image Representation As radiology lies outside the distribution of most general-domain image models, a radiology-specific pre-trained model is crucial for the RRG task (Pérez-García et al., 2024).

Zhou et al. (2023) introduces a unified framework, Masked Record Modeling (MRM), which combines self-supervision and radiology report supervision to enhance radiograph representation learning. BioViL-T (Bannur et al., 2023) integrates a CNN-Transformer hybrid architecture for multi-modal modelling, leveraging temporal structures in medical imaging and reports to perform tasks such as disease progression classification and report generation. BiomedCLIP (Zhang et al., 2024b) is a multimodal biomedical foundational model pre-trained on various biomedical tasks. RAD-DINO (Pérez-García et al., 2024) is a medical image encoder that employs a pure image-based self-supervised learning approach from DINOv2 (Oquab et al., 2024) for continuous pretraining, focusing exclusively on image data to avoid the limitations of text supervision.

In recent parallel works, RAD-DINO has been widely applied to RRG tasks, including MAIRA-2 (Bannur et al., 2024), M4CXR (Park et al., 2024). Furthermore, Pérez-García et al. (2024) highlighted that RAD-DINO outperforms other types of image encoders in RRG tasks. Therefore, our model also adopts RAD-DINO as the image encoder.

A.2 Research Objectives

Temporal Information Temporal change is a crucial factor in understanding disease progression, especially in radiology, where paired images and corresponding radiology reports capture changes over time. This temporal context is often reflected in comparisons between current and prior scans, documenting symptom evolution or newly identified findings.

The relativity of position within the timeline is a key determinant of temporal information. This relative positioning does not alter the established description of well-established facts about symptoms already present in the current image. For instance, if the prior image is recent, descriptions of changes will be minimal, whereas an older prior image may highlight more pronounced changes.

Research Object Our study focuses on the chest X-ray (frontal view), rather than the individual patient. Specifically, we aim to generate a report for the current image while referencing the temporal information provided by a single prior image as auxiliary information.

Our model is primarily designed to handle scenarios with limited temporal information in the RRG task. For instance, in a case where a patient has two scans taken just milliseconds apart, the current and prior images would be nearly identical, as no pathological changes would manifest within such a short interval. This extreme scenario demonstrates how the model handles clinical practice under limited temporal information. In such cases, the correct diagnosis for this minimal interval would be that the patient's condition is "stable." In our design, the dummy prior image addresses scenarios where a true prior image is absent.

However, patients often undergo multiple prior scans, which may include temporal information from different orientations. This more complex temporal information lies beyond the scope of our current study. We will provide a detailed discussion of this scenario in Appendix A.3.

A.3 Limitations and Future Work

While our work represents a step forward in leveraging temporal information for radiology report generation, it also has several limitations that warrant further exploration.

Handling Multiple Prior Scans Our current model is designed to process a single prior scan alongside the current scan. While this approach aligns with standard clinical workflows, which typically

prioritise the most recent prior study for comparisons, it overlooks scenarios where multiple prior scans could offer a richer temporal perspective. For instance, analysing a sequence of images spanning an extended period could provide deeper insights into gradual disease progression. Future efforts should focus on extending our framework to incorporate multiple prior scans efficiently, enabling a more nuanced understanding of temporal patterns in clinical data.

Temporal Information Beyond Image Comparisons At present, our model captures temporal information through paired image comparisons and their corresponding textual reports. However, clinical assessments often draw upon a broader context, including historical notes, laboratory results, and other longitudinal patient data. Expanding our approach to integrate these diverse temporal data sources could facilitate a more holistic understanding of disease trajectories and patient history, significantly enhancing clinical applicability.

Challenges with Sparse Temporal Data In cases where prior scans are unavailable or only minimally informative (e.g., scans taken within a very short time interval), our use of a "dummy prior image" addresses the lack of temporal data. However, the model's ability to interpret and generate meaningful outputs under these constraints may still be limited. Future work could explore methods for imputing or synthesising temporal context to address such cases more effectively.

Computational Complexity Our use of temporal alignment mechanisms and multi-layer feature integration introduces increased computational demands. This complexity poses challenges for real-time deployment, particularly in resource-constrained clinical environments.

Generalisability Across Modalities and Datasets This study focuses exclusively on frontal-view chest X-rays and utilises the MIMIC-CXR dataset (Johnson et al., 2019b). However, our approach has not been evaluated on other imaging modalities (e.g., CT or MRI) or datasets such as CheXpert (Irvin et al., 2019b) or PadChest (Bustos et al., 2020).

Based on the identified limitations, we propose several directions for future research and development:

- Develop frameworks for integrating multiple prior scans with dynamic temporal reasoning.
- Extend the model to handle multi-modal imaging and textual data for comprehensive diagnostics.
- Investigating the integration of temporal data from diverse clinical sources, including electronic health records (EHRs)
- Exploring lightweight model architectures for faster inference while maintaining high performance.

These enhancements will not only address the current limitations but also extend the potential of temporal-aware multimodal models in radiology and beyond.

B Appendix B

B.1 Training Configuration

We train the model using a standard auto-regressive language modelling loss (cross-entropy). For this study, we utilise Meditron-7b (Chen et al., 2023b) as the LLM, with a total batch size of 16 throughout the training process. The training is conducted on a computational infrastructure equipped with A6000 GPU (48GB of memory) and configured with DeepSpeed ZeRO-3 optimization (Rajbhandari et al., 2020) and BF16 precision enabled. A cosine learning rate scheduler is employed, starting with a warm-up phase of 0.03. In the first stage of training, we run for 1 epoch with a learning rate of 2×10^{-5}. In the second stage, the model is trained for 3 epochs at the same learning rate. The LoRA (Hu et al., 2021) parameters are set to $r = 128$ and $alpha = 256$. The final checkpoint for all runs is selected based on the observation of the minimum loss on the evaluation dataset throughout the training process.

B.2 Detailed Description of Datasets

MIMIC-CXR. (Johnson et al., 2019b) This is a large, publicly accessible dataset comprising 377,110 DICOM images across 227,835 studies, each accompanied by a radiology report (John-

son et al., 2019b). For images, we use the commonly available JPEG files from MIMIC-CXR-JPG (Johnson et al., 2019a), rather than the original DICOM files, and we preprocess the dataset to exclude non-AP/PA scans. For each report, we extract the *Findings, Impression, Indication, History, Comparison,* and *Technique* sections using rule-based heuristics supported by the official MIMIC code repository (Johnson et al., 2018).

For the *Findings* section generation task, we discard studies where the *Findings* could not be extracted, while allowing for missing content in other clinical sections. The same approach is applied to the *Impression* section generation task. In all our experiments, we adhere to the official MIMIC-CXR dataset split. Additionally, we retrieve prior images by following the chronological order of studies as indicated by the official labels, selecting the closest prior study as the reference image. It is important to note that, to prevent data leakage between the train, validation, and test sets, prior images are retrieved only from within the same split.

Medical-Diff-VQA. (Hu et al., 2023) This dataset is a derivative of the MIMIC-CXR dataset, focused on identifying differences between pairs of main and reference images. The data split adheres to the original labelling, ensuring no data leakage occurs. In total, this dataset comprises 700,703 question-answer pairs derived from 164,324 main-reference image pairs. The questions are divided into seven categories: abnormality, location, type, view, presence, and difference. Each pair consists of a main (current) image and a reference (prior) image, both taken from different studies of the same patient. The reference image is always selected from an earlier visit, with the main image representing the later visit. Of the seven question types, the first six types focus on the main image, while the "difference" questions involve both images.

MIMIC-Ext-*MIMIC-CXR-VQA*. (Bae et al., 2023) This dataset is an extension of the MIMIC-CXR dataset, specifically designed for VQA tasks within the medical domain, specifically focused on CXRs. It includes questions generated from 48 unique templates covering seven content types: presence, anatomy, attribute, abnormality, size, plane, and gender. Each template was developed with the guidance of board-certified medical experts to ensure clinical relevance, addressing both standard medical VQA content and more complex logical scenarios. In total, the dataset consists of 377,391 unique entries. Since annotations are based on single images, the current image serves as a dummy prior image for all entries in our experiment.

For this study, we carefully selected datasets that provide complete reports and temporal information (i.e., prior images) to ensure alignment with our research objectives for the RRG task. After thoroughly evaluating other datasets, we found them unsuitable for the following reasons:

CheXpert. (Irvin et al., 2019b) This dataset includes annotated scans with label-specific annotations rather than full medical reports. While useful for training image encoders or annotation models, it is not appropriate for the RRG task, which requires complete diagnostic reports.

PadChest. (Bustos et al., 2020) Although it includes reports and corresponding prior images, its reports are in Spanish, placing cross-language training beyond the scope of our model.

IU-Xray. (Demner-Fushman et al., 2016) This dataset lacks patient-level metadata and prior study information, which is critical for our focus on temporal information in chest X-rays.

Chest ImaGenome Dataset. (Wu et al., 2021a)(a derivative of MIMIC-CXR) This dataset does not follow the official MIMIC-CXR split, risking data leakage across train, validation, and test sets.

B.3 Evaluation Metrics Details

Lexical Metrics We employed common natural language generation metrics to quantify the overlap between generated and reference reports. Specifically, ROUGE-L (Lin, 2004) measures the length of the longest common subsequence between the generated and reference reports. BLEU-$\{1, 4\}$ (Papineni et al., 2002) calculates n-gram precision and applies a brevity penalty to discourage overly short predictions. METEOR (Banerjee & Lavie, 2005), computes the weighted harmonic mean of unigram precision and recall, with an additional penalty for fragmenting consecutive word sequences. Finally, we report BERTScore (Zhang et al., 2020a), which leverages pre-trained contextual embeddings from BERT (Devlin et al., 2019) to match words in candidate and reference sentences based on cosine similarity. We used default parameters for all of these evaluation metrics.

For radiology-specific metrics, we used as many of the same evaluation scores as possible from previous studies (Tu et al., 2023; Hyland et al., 2024; Bannur et al., 2024; Chaves et al., 2024), including the following:

RadGraph-based metrics The RadGraph model (Jain et al., 2021) is designed to parse radiology reports into structured graphs. These graphs consist of clinical entities, which include references to anatomy and observations, as well as the relationships between these entities. This structured representation enables a more detailed and systematic analysis of radiology reports, facilitating downstream tasks such as information extraction, report generation, and clinical decision support.

These include RadGraph-F1 (Jain et al., 2021), which computes the overlap in entities and relations separately and then reports their average. And a variant of it, RG_{ER} (Delbrouck et al., 2022b), which matches entities based on their text, type, and whether they have at least one relation[6].

CheXpert F1 This set of metrics utilizes the CheXbert automatic labeler (Smit et al., 2020) to extract "present", "absent", or "uncertain" labels for each of the 14 CheXpert pathologies (Irvin et al., 2019a) from the generated reports and their corresponding references. In line with prior work, we report CheXpert-F1 for all 14 classes, as well as for the 5 most common findings in real-world CXR reports, referring to these as '[Macro/Micro]-F1-[5/14]'.

CheXbert vector similarity We also employ CheXbert vector similarity (Yu et al., 2022), which calculates the cosine similarity between the embeddings of the generated and reference reports after processing them through the CheXbert model (Smit et al., 2020).

RadCliQ In addition, we utilise RadCliQ (Radiology Report Clinical Quality), a composite metric that combines RadGraph-F1 and BLEU scores in a linear regression model to estimate the number of errors that radiologists are likely to detect in a report. To maintain consistency with previous research, we use version 0 of it.

Both the CheXbert vector similarity, RadCliQ_0, and RadGraph-F1 metrics are calculated using the code released by Yu et al. (2022).

Temporal Entity F1 The $F1_{\text{temp}}$, which is specifically designed to detect temporal entities. This score is neither a traditional lexical metric nor a adiology-specific metrics. Here is an example to help you understand better, as shown in Table 6.

Table 6: Evaluation of Candidate Using Temporal Entity F1

Ground Truth	**Candidate**	ROUGE-L	RadGraph-F1	$F1_{\text{temp}}$
Compare with prior scan, pleural effusion has worsened.	The pleural effusion has progressively worsened since previous scan.	0.47	0.86	**1.0**
	The pleural effusion is noted again on the current scan.	0.22	0.80	**0.0**

From these results, it is clear that the differences in lexical (ROUGE-L) and clinical (RadGraph-F1) metrics between the two candidates are relatively smaller compared to the $F1_{\text{temp}}$ score. This highlights that Temporal Entity F1 effectively captures and evaluates the quality of temporal information in radiology reports, distinguishing it more accurately than other standard metrics in the context of temporal information descriptions.

[6] RG_{ER} is implemented as `F1RadGraph` with `reward=partial` by the radgraph package.

B.4 PROMPT EXAMPLE

Here, we select examples from the MIMIC-CXR (Johnson et al., 2019b) dataset as instances and further synthesise them using GPT-4 (OpenAI et al., 2024) to avoid any ethical concerns, as shown in Table 7. Following Hyland et al. (2024)'s work, we adopt a rule-based method to extract relevant sections from the report of the current image, utilising a fixed system prompt and a dynamic clinical prompt for each current scan. We utilised four clinical instructions from the original report: {*Indication*}, {*History*}, {*Comparison*}, and {*Technique*}. Notably, unlike MAIRA-2 (Bannur et al., 2024), we do not utilise the report of the prior image.

Table 7: Examples of Prompt for Libra in RRG task.

Original Radiology Report
EXAMINATION: Chest (Portable AP) INDICATION: Dyspnea and cough, right-sided back pain. HISTORY: Intubation with pulmonary edema. COMPARISON: Chest radiographs on ___ and CT chest without contrast on ___. TECHNIQUE: Portable upright chest radiograph. FINDINGS: In comparison with the prior study, there are diffuse bilateral pulmonary opacifications, more prominent on the right. These findings could indicate severe pulmonary edema, but superimposed pneumonia or developing ARDS cannot be excluded. Monitoring and support devices are appropriately positioned.
Prompt Content

[System prompt] {
The assistant specialised in comparing Chest X-ray images, identifying differences, and noting temporal changes.
}
+
<Image Representation Placeholder>
+
[Clinical prompt] {
Provide a detailed description of the findings in the radiology image. Following clinical context:
Indication: Dyspnea and cough, right-sided back pain.
History: Intubation with pulmonary edema.
Comparison: Chest radiographs on ___ and CT chest without contrast on ___.
Technique: Portable upright chest radiograph.
}

C Appendix C

C.1 Impact of the Temporal Alignment Connector under General-Domain Pre-trained Models

Table 8: Results of ablation experiments for the Temporal Alignment Connector. '↓' indicates that lower is better. Values in (%) indicate the percentage decrease compared with the Libra-b.

Metric	Libra-b	w/o TFM	w/o LFE	w/o PIPB	w/o TAC
Lexical:					
ROUGE-L	27.26	26.80 (-1.69%)	26.57 (-2.53%)	27.00 (-0.95%)	24.58 (-9.83%)
BLEU-1	34.94	33.61 (-3.81%)	33.68 (-3.61%)	34.47 (-1.35%)	31.40 (-10.13%)
BLEU-4	11.74	10.97 (-6.56%)	10.94 (-6.81%)	11.57 (-1.45%)	8.89 (-24.28%)
METEOR	35.37	34.50 (-2.46%)	34.30 (-3.03%)	34.93 (-1.24%)	32.41 (-8.37%)
BERTScore	55.51	54.97 (-0.97%)	54.75 (-1.37%)	55.26 (-0.45%)	53.07 (-4.40%)
$F1_{temp}$	24.77	23.54 (-4.97%)	24.00 (-3.11%)	24.68 (-0.36%)	22.52 (-9.08%)
Clinical:					
RadGraph-F1	21.67	21.06 (-2.81%)	20.73 (-4.34%)	21.35 (-1.48%)	19.77 (-8.77%)
RG_{ER}	26.28	25.45 (-3.16%)	25.14 (-4.34%)	25.84 (-1.67%)	23.74 (-9.67%)
$RadCliQ_0$ (↓)	3.17	3.20 (-0.95%)	3.22 (-1.58%)	3.18 (-0.32%)	3.27 (-3.15%)
CheXbert vector	39.58	38.74 (-2.12%)	38.37 (-3.06%)	39.49 (-0.23%)	37.56 (-5.10%)
CheXpert-F1:					
Micro-F1-14	49.06	47.68 (-2.81%)	47.57 (-3.04%)	48.40 (-1.35%)	46.57 (-5.08%)
Macro-F1-14	33.07	31.60 (-4.45%)	31.78 (-3.90%)	31.78 (-3.90%)	31.27 (-5.44%)
Micro-F1-5	54.55	52.14 (-4.42%)	52.94 (-2.95%)	53.10 (-2.66%)	50.72 (-7.02%)
Macro-F1-5	47.24	44.28 (-6.27%)	44.39 (-6.04%)	44.68 (-5.42%)	43.48 (-7.96%)

Domain-specific pre-trained models (i.e., RAD-DINO (Pérez-García et al., 2024) and Meditron (Chen et al., 2023b)) inherently incorporate domain-specific knowledge, such as phrasing conventions, pronoun usage, and even temporal information embedded in the training corpus.

To eliminate the influence of such factors, we used a general-domain image encoder (DINOv2 (Oquab et al., 2024)) and a LLM (Vicuna-7B-v1.5 (Chiang et al., 2023)), allowing the structural enhancements of TAC to be observed more directly. We adopted the same setup as the second ablation study in Section 4. We first conducted a baseline experiment, referred to as Libra-b, by fine-tuning only the adapter for the *Findings* generation task. As shown in Table 8, we then conducted ablation studies by sequentially removing different components from the model, including the TFM, LFE, Prior Image Prefix Bias (PIPB) in TFM, and the entire TAC. Removing TFM restricts the model to processing only the current image, using a configuration similar to LLaVA (Liu et al., 2023), but with a four-layer MLP to align the image representation with the LLM's hidden dimensions. It is important to note that without the TFM module, the model cannot process true prior images or dummy prior images, and is limited to only the current image as input. Without LFE, the model follows the LLaVA setup, using the penultimate layer of the image encoder to process single or paired images.

The ablation results are consistent with those observed when using domain-specific pre-trained models, as presented in Section 4. Compared to Libra-b, removing any TAC submodule results in a decline across all metrics. Removing the TFM caused a notable drop in the $F1_{temp}$ score (↓>4%), emphasising its role in capturing temporal information. The removal of the LFE significantly decreased RadGraph-related scores, highlighting its importance for extracting detailed image features. Removing the PIPB impacted clinical scores more than lexical scores. Finally, removing the TAC resulted in significant declines across all metrics, underscoring its critical role in integrating image details and temporal information.

C.2 Impact of Radiology-Specific Pre-trained Models

Aligning radiology images with textual information is a key challenge in RRG tasks. To demonstrate the benefits of using radiology-specific pre-trained models for more accurate feature representation and improved MLLM performance, we initialised a Libra model with RadDINO, the TAC, and Meditron-7b, conducting the first stage of training, denoted as Libra-1 (This is consistent with the

Table 9: Ablation results for radiology-specific pre-trained models in Libra. '↓' indicates that lower is better. Values in (%) indicate the percentage improvement compared to Libra-c.

Metric	Libra-1	w/o RadDINO	w/o Meditron	w/o RadDINO+Meditron
Lexical:				
ROUGE-L	27.56	27.66 (0.36%)	27.29 (-0.98%)	27.26 (-1.09%)
BLEU-1	34.84	35.32 (1.38%)	34.91 (0.20%)	34.94 (0.29%)
BLEU-4	11.51	12.56 (9.12%)	11.61 (0.87%)	11.74 (2.00%)
METEOR	35.50	35.65 (0.42%)	35.53 (0.08%)	35.37 (-0.37%)
BERTScore	55.87	55.89 (0.04%)	55.58 (-0.52%)	55.51 (-0.64%)
$F1_{temp}$	26.63	25.53 (-4.13%)	24.78 (-6.95%)	24.77 (-6.98%)
Clinical:				
RadGraph-F1	22.52	22.11 (-1.82%)	23.13 (2.71%)	21.67 (-3.77%)
RG_{ER}	27.32	26.72 (-2.20%)	27.53 (0.77%)	26.28 (-3.81%)
$RadCliQ_0$ (↓)	3.10	3.13 (-0.97%)	3.08 (0.65%)	3.17 (-2.26%)
CheXbert vector	42.02	40.78 (-2.95%)	41.94 (-0.19%)	39.49 (-6.02%)
CheXpert-F1:				
Micro-F1-14	52.84	51.55 (-2.44%)	51.45 (-2.63%)	49.06 (-7.15%)
Macro-F1-14	36.87	34.58 (-6.21%)	37.20 (0.90%)	33.07 (-10.31%)
Micro-F1-5	56.63	55.00 (-2.88%)	55.39 (-2.19%)	54.55 (-3.67%)
Macro-F1-5	49.33	47.26 (-4.20%)	47.62 (-3.47%)	47.24 (-4.24%)

baseline setup of the second ablation study in Section 4). Then we replaced the image encoder and LLM with their general-domain counterparts, DINOv2 and Vicuna-7B-v1.5, respectively. Finally, we replaced both components, which is also referred to as Libra-b.

As shown in Table 9, substituting radiology-specific pre-trained models with general-domain models resulted in a notable decline in clinical scores, while the impact on lexical scores was minimal. Notably, replacing the radiology-specific image encoder caused a more pronounced decline in clinical metrics compared to replacing the language model. This suggests that accurate medical image representation provides greater benefits in RRG tasks, indicating the importance of incorporating domain-specific knowledge into pre-trained models to enhance Libra's performance.

C.3 Incremental Study

In this section, we conducted an incremental study of our model, mirroring the logical framework we used to construct Libra. We started with an architecture similar to LLaVA, comprising a pre-trained CLIP image encoder, a randomly initialised four-layer MLP as the adapter, and Vicuna-7B-v1.5 as the LLM. We train the adapter on the *Findings* section generation task to establish an initial baseline.

Gradual enhancements are then introduced, as shown in the Table 10. First, the image encoder is replaced with DINOv2. Building upon this, we integrate the LFE, the prefix module of the TAC connector. Subsequently, we add the TFM, the suffix module, thus completing the full TAC connector. Following this, the image encoder and LLM are progressively replaced with biomedical domain models, RAD-DINO and Meditron, respectively. The dataset used for pretraining in the first stage is then extended. Finally, the model is fine-tuned for downstream tasks to obtain Libra.

We observed that with each step of improvement, the overall performance consistently exceeded the previous state, demonstrating the essential role of each component in our optimisation scheme. Notably, in the alignment stage, the model showed the greatest overall improvement after the addition of the TFM, highlighting its capability to capture temporal information, which significantly enhanced performance on the RRA task.

However, after the data expansion optimisation in the first stage, the scores on lexical metrics improved, while clinical metrics exhibited a slight decline. This can be attributed to the introduction of the VQA task, which led the model to perceive more fine-grained grounded information. This is distinct from the report generation task, which focuses on describing the holistic level. Furthermore, this shift also affects the $F1_{temp}$ score, as temporal entities are often associated with the presence of specific symptoms. Ultimately, these dropped scores are improved again through the second stage of fine-tuning.

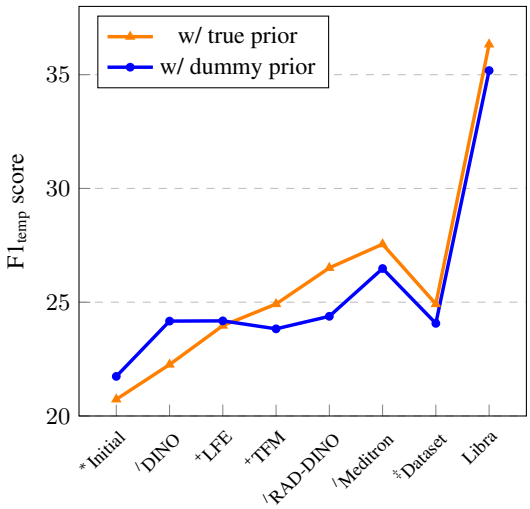
Table 10: Results of ablation experiments for key components of Libra on *Findings* section generation performance. * indicates our initialised model. / denotes component replacement. + signifies structural addition. ‡ represents dataset expansion. The best performances are highlighted in **bold**, and the second-best scores are underlined. '↓' indicates that lower is better.

Metric	Stage 1: Temporal Feature Alignment							Stage 2
	*Initial	/DINO	+LFE	+TFM	/RAD-DINO	/Meditron	‡Dataset	Libra
Lexical:								
ROUGE-L	23.77	24.58	26.57	27.26	27.29	27.56	27.27	**36.66**
BLEU-1	31.48	31.40	33.68	34.94	34.91	34.84	41.24	**51.25**
BLEU-4	8.41	8.89	10.94	11.74	11.61	11.51	13.59	**24.54**
METEOR	32.1	32.41	34.3	35.37	35.53	35.50	39.44	**48.90**
BERTScore	52.76	53.07	54.75	55.51	55.58	55.87	56.00	**62.50**
$F1_{temp}$	21.60	22.52	24.00	24.77	24.78	26.63	24.80	**35.34**
Clinical:								
RadGraph-F1	18.58	19.70	20.73	21.67	23.13	22.52	20.45	**32.87**
RG_{ER}	23.05	23.74	25.14	26.28	27.53	27.32	25.19	**37.57**
$RadCliQ_0$(↓)	3.35	3.26	3.22	3.17	3.08	3.10	3.31	**2.72**
CheXbert vector	35.59	37.94	38.37	39.49	41.94	42.02	35.33	**46.85**
CheXpert-F1:								
Micro-F1-14	44.75	46.57	47.57	49.06	51.45	52.48	43.63	**55.87**
Macro-F1-14	25.13	31.27	31.07	33.07	37.20	36.87	25.68	**40.38**
Micro-F1-5	45.97	50.72	52.94	54.55	55.39	56.63	49.75	**60.07**
Macro-F1-5	36.55	43.48	44.39	47.24	47.62	49.33	40.40	**53.75**

Evaluation of Libra's Temporal Awareness. Another approach to investigating the model's ability to capture temporal information is to evaluate it separately within the test split based on the presence or absence of true prior images, as shown in Table 11.

With the addition of the TFM, the model exhibited temporal awareness. It is worth noting that, for the first time, the $F1_{temp}$ score of samples with prior images surpassed those without, and this trend persisted through subsequent optimisations. This indicates that the structural enhancements have resulted in a sustained improvement in the model's temporal perception capabilities. An effective example is in Section 5.

Table 11: Results of ablation experiments for Libra on the $F1_{temp}$ score. Of the 2,461 official test samples, 2,117 include a prior image as a reference and 344 do not.



D APPENDIX D

D.1 DISCUSSION ON PERFORMANCE WITH RADIOLOGY FOUNDATION MODELS

Table 12: Findings Generation Performance of Libra and the Latest Concurrent Work. The best performances are highlighted in **bold**, and the second-best scores are underlined. '↓' indicates that lower values are better. The percentage (%) indicates the improvement over the best existing model.

Metric	MedVersa	MAIRA-2	M4CXR	Libra (%)
Lexical:				
ROUGE-L	–	**38.4**	28.5	36.7 (-4.4%)
BLEU-1	–	46.5	33.9	**51.3** (10.3%)
BLEU-4	17.8	23.4	10.3	**24.5** (4.7%)
METEOR	–	42.0	–	**48.9** (16.4%)
BERTScore	49.7	–	–	**62.5** (25.8%)
Clinical:				
RadGraph-F1	28.0	**34.6**	21.8	32.9 (-4.9%)
RG_{ER}	–	**39.7**	–	37.6 (-5.3%)
$RadCliQ_0$(↓)	2.7	**2.6**	–	2.7 (-3.8%)
CheXbert vector	46.4	**50.6**	–	46.9 (-7.3%)
CheXpert-F1:				
Micro-F1-14	–	58.5	**60.6**	55.9 (-7.8%)
Macro-F1-14	–	**42.7**	40.0	40.4 (-5.4%)
Micro-F1-5	–	58.9	**61.8**	60.1 (-2.8%)
Macro-F1-5	–	51.5	49.5	**53.8** (4.5%)

As shown in Table 12, these models belong to the category of radiology foundation models. MedVersa (Zhou et al., 2024) and M4CXR (Park et al., 2024) support multiple tasks, including medical report generation, visual grounding, and visual question answering. In contrast, MAIRA-2 (Bannur et al., 2024) specialises in grounded radiology report generation, a novel task requiring image-level localisation for each identifiable finding or symptom. It is worth noting that the test sets differ slightly. All of these models utilise additional radiology information, such as lateral views, prior study reports of CXRs, or both.

Despite these considerations, Libra achieves the highest scores on most lexical metrics, including BLEU-$\{1, 4\}$, METEOR, and BERTScore, while trailing slightly behind MAIRA-2 on ROUGE-L. In clinical metrics, Libra predominantly ranks second, just behind the best-performing model. For clinical metrics, Libra consistently ranks second, just behind the top-performing model. In metrics that evaluate medical entities and their relationships, such as RadGraph-F1, RG_{ER}, and RadCliQ, Libra also ranks second. Similarly, Libra comes second in the CheXbert vector embedding score. However, in the CheXpert metrics, Libra ranks first in Macro-F1 for the 5-class subset, with only a slight dip in the Micro-F1 score for the 14-class subset.

Incorporating lateral images and prior study reports could enhance clinical scores. Additionally, strategies like chain-of-thought reasoning and grounded report generation further improve performance in RRG tasks. Looking ahead, we plan to develop model architectures that can automatically adapt to multiple tasks and diverse scenarios, enabling more efficient handling of additional radiological information.

D.2 Discussion on Performance with non-LLM-based Models

Table 13: Findings Generation Performance of Libra and non-LLM-based Models. The best performances are highlighted in **bold**, and the second-best scores are underlined. These results are taken from the best performances reported in their original papers.

Model	Lexical Metrics						Clinical Metrics		
	B-1	B-2	B-3	B-4	MTR	R-L	P	R	F_1
R2Gen	35.3	21.8	14.5	10.3	14.2	27.0	33.3	27.3	27.6
R2GenCMN	35.3	21.8	14.8	10.6	14.2	27.8	34.4	27.5	27.8
M^2TR	37.8	23.2	15.4	10.7	14.5	27.2	24.0	42.8	30.8
KnowMAT	36.3	22.8	15.6	11.5	–	28.4	45.8	34.8	37.1
CMM-RL	38.1	23.2	15.5	10.9	15.1	28.7	34.2	29.4	29.2
CMCA	36.0	22.7	15.6	11.7	14.8	28.7	44.4	29.7	35.6
KiUT	39.3	24.3	15.9	11.3	16.0	28.5	37.1	31.8	32.1
DCL	–	–	–	10.9	15.0	28.4	47.1	35.2	37.3
METrans	25.0	16.9	12.4	15.2	–	29.1	36.4	30.9	31.1
ORGAN	38.6	25.6	17.2	12.3	16.2	29.3	41.6	41.8	38.5
COMG	36.3	23.5	16.7	12.4	12.8	29.0	–	–	–
MedM2G	41.2	26.9	17.9	14.2	–	30.9	–	–	–
BioViL-T	–	–	–	9.2	–	29.6	–	–	17.5
TiBiX	32.4	23.4	18.5	15.7	16.2	33.1	30.0	22.4	25.0
RECAP	42.9	26.7	17.7	12.5	16.8	28.8	38.9	44.3	39.3
Libra	**51.3**	**38.0**	**30.0**	**24.5**	**48.9**	**36.7**	**59.7**	**52.5**	**55.9**

To facilitate comparison with non-LLM-based models, we selected evaluation metrics commonly used in these studies. These include BLEU-$\{1, 2, 3, 4\}$ (Papineni et al., 2002), METEOR (MTR) (Banerjee & Lavie, 2005), and ROUGE-L (R-L) (Lin, 2004). For clinical metrics, we adopted the CheXbert (Irvin et al., 2019a), reporting Precision (P), Recall (R), and F_1.

Baseline For performance evaluation, we compare our model with the following baselines: R2Gen (Chen et al., 2021), R2GenCMN (Chen et al., 2021), M^2TR (Nooralahzadeh et al., 2021), KnowMAT (Yang et al., 2022), CMM-RL (Qin & Song, 2022), CMCA (Song et al., 2022), KiUT (Huang et al., 2023), DCL (Li et al., 2023b), METrans (Wang et al., 2023b), ORGAN (Hou et al., 2023b), COMG (Gu et al., 2023), MedM2G (Zhan et al., 2024), BioViL-T (Bannur et al., 2023), TiBiX (Sanjeev et al., 2024), RECAP (Hou et al., 2023a).

As the results demonstrate, our model, like other LLM-based models, generally outperforms non-LLM-based models. This is attributed to advancements in LLMs and visual instruction tuning (Liu et al., 2023), enabling multimodal large language models (MLLMs) to excel in the RRG task.

E APPENDIX E

E.1 TEMPORAL ALIGNMENT HEATMAP ANALYSIS AND FEATURE REPRESENTATION

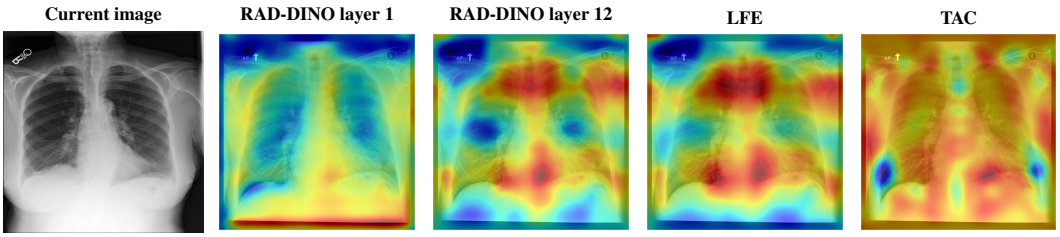


Figure 4: Heat map visualisation of image representations from different image encoder layers and the Temporal Alignment Connector (TAC), up-sampled using a Gaussian filter. Warm colours (red, yellow) indicate regions with higher weight allocations in the intermediate outputs of the "hidden-state" within the model blocks, while cool colours (blue, green) represent regions with lower weight.

The heatmap in Figure 4 corresponds to the example in Figure 2, where no true prior image was used as a reference. It highlights the clear differences in feature representations across layers of the RAD-DINO image encoder. The shallow layers primarily capture the overall lung structure, while the deeper layers focus on specific disease locations.

After being processed by the Layerwise Feature Extractor (LFE), the image feature representations show that larger symptom regions are assigned higher weights. This indicates that the image representation achieves a more fine-grained level of detail. Following the Temporal Alignment Connector (TAC), Libra integrates the weighted dummy prior image, resulting in a more uniform feature distribution. This reflects temporal information, indicating no significant changes compared to the prior study. This process helps the model provide a smoother image feature representation for subsequent processing by the text decoder (i.e., the LLM).

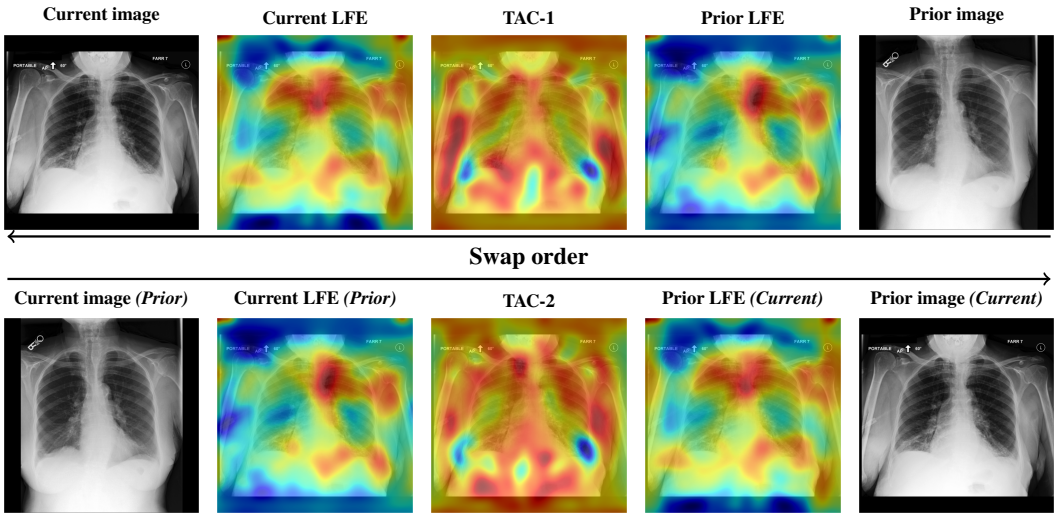


Figure 5: Heat map visualisation of image representations from the Temporal Alignment Connector (TAC), up-sampled using a Gaussian filter. The arrows ('→') represent the direction of temporal information, pointing from the true prior image to the true current image. Warm colours (red, yellow) indicate regions with higher weight allocations in the intermediate outputs of the "hidden-state" within the model blocks, while cool colours (blue, green) represent regions with lower weight.

The heatmap in Figure 5 corresponds to the example in Figure 3, where a true prior image is provided. As observed, after processing through the LFE, the model captures higher-granularity feature

representations in the symptom areas. After being processed by the TAC, the intermediate representation not only integrates these fine-grained features but also incorporates the difference information between the two images, reflecting the temporal information provided by the reference image, as shown in TAC-1 (Figure 5).

When the image order is swapped, treating the true prior image as the current image, the intermediate output from the LFE remains unchanged. Nevertheless, a comparison of the TAC-2 and TAC-1 outputs reveals noticeable differences in the lung feature representations. This demonstrates that the model's temporal perception is directional and that the TAC module successfully integrates temporal information from different time points. At the same time, it confirms that the LFE focuses solely on extracting image feature representations without encoding temporal aspects.

This result is expected because the residual connections in the TAC ensure that the current image remains the main modality, while the prior image serves as the auxiliary modality. When the order is swapped, the prior image becomes the main modality for subsequent processing. Thus, swapping the image order also changes the content of the generated report, as mentioned in Section 5. The primary difference lies in the reversal of the temporal state of symptoms, such as changing from "improving" to "worsening."