# Libra: Leveraging Temporal Images for Biomedical Radiology Analysis

**Anonymous ACL submission**

## Abstract

Radiology report generation (RRG) requires advanced medical image analysis, effective temporal reasoning, and accurate text generation. While multimodal large language models (MLLMs) align with pre-trained vision encoders to enhance visual-language understanding, most existing methods rely on single-image analysis or rule-based heuristics to process multiple images, failing to fully leverage temporal information in multi-modal medical datasets. In this paper, we introduce **Libra**, a temporal-aware MLLM tailored for chest X-ray report generation. Libra combines a radiology-specific image encoder with a novel Temporal Alignment Connector (**TAC**), designed to accurately capture and integrate temporal differences between paired current and prior images. Extensive experiments on the MIMIC-CXR dataset demonstrate that Libra establishes a new state-of-the-art benchmark among similarly scaled MLLMs, setting new standards in both clinical relevance and lexical accuracy.

## 1 Introduction

Radiology reports are critical for biomedical radiology analysis, offering structured summaries of imaging studies such as chest X-rays (CXRs). Commonly divided into sections like *Findings, Impression, Indication, Technique, Comparison*, and *History* (Ganeshan et al., 2018), these reports guide diagnostic and therapeutic decisions (Najjar, 2023). However, manually generating such reports is both complex and time-consuming. Automating radiology report generation (RRG) holds great promise for alleviating radiologist burnout, increasing efficiency, and improving communication (Zhang et al., 2020b). Despite this, the intricate nature of medical imaging demands precise and detailed documentation, making RRG a challenging task.

Recent advances in Multimodal Large Language Models (MLLMs), such as LLaVA (Liu et al., 2023) and InstructBLIP (Dai et al., 2023), have demonstrated potential in vision-language tasks. However, their performance diminishes in specialised biomedical contexts due to the significant domain shift between general-purpose and medical image-text data (Tu et al., 2023). These models often lack fine-grained detail for medical imaging tasks, resulting in surface-level understanding akin to layperson interpretations. While continued pre-training on medical datasets and domain-specific fine-tuning (e.g., LLaVA-Med (Li et al., 2023a), MedBLIP (Chen et al., 2023b)) improve performance, they still cannot fully capture the complexities of medical image analysis (Xiao et al., 2024).

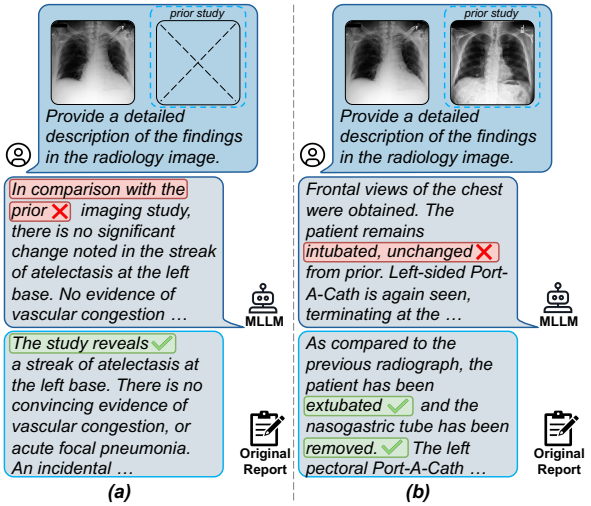


Figure 1: Examples of hallucinations in RRG using the MLLM (MAIRA-1 (Hyland et al., 2024)). ***(a)*** Single-image case: spurious references to nonexistent prior studies. ***(b)*** Temporal image case: inaccurate interpretation of temporal changes when integrating prior studies.

One critical gap in the current MLLM-based approaches is their limited ability to incorporate temporal context, which is pivotal in clinical practice. Radiologists routinely compare current imaging results with prior studies to identify temporal changes, a process crucial for understanding disease progression and guiding treatment decisions. Indeed, the MIMIC-CXR database (Johnson et al., 2019b) reveals that **67%** of patients underwent at

least two studies at different time intervals, underscoring the necessity of temporal reasoning. However, most MLLMs designed for RRG tasks focus on single-image analysis, neglecting this temporal dimension (Zhang et al., 2024c). As illustrated in Figure 1, MAIRA-1 (Hyland et al., 2024) introduces hallucinated prior references in single-image analysis and misinterprets temporal changes when integrating prior studies.

Although recent models[1] like MedVersa (Zhou et al., 2024) and MAIRA-2 (Bannur et al., 2024) have introduced multi-image processing, they do not explicitly model or extract temporal differences. Instead, they rely on inserting visual tokens from different studies at specific points within textual inputs and delegate the reasoning task to the LLM. Similarly, Banerjee et al. (2024) and Chaves et al. (2024) leverage GPT-4V (OpenAI et al., 2024) to eliminate hallucinated references to prior studies in the dataset but lacks dedicated mechanisms for modelling temporal progression. Additionally, existing MLLMs often rely on embeddings from the last or penultimate layer of the image encoder (Chen et al., 2023a; Zhang et al., 2024a), primarily capturing global features. However, RRG tasks require fine-grained details[2] (Sloan et al., 2024), which a single-layer embedding often cannot fully represent (Jiang et al., 2024). To tackle these limitations, we enhance MLLM temporal awareness for RRG tasks by addressing two main challenges:

- Designing robust MLLM architectures that seamlessly handle prior study references in RRG.
- The scarcity of effective feature alignment projectors in MLLMs capable of handling the high-granularity requirements of downstream tasks.

To overcome these gaps, we propose **Libra** (**L**everaging Temporal **I**mages for **B**iomedical **R**adiology **A**nalysis), a novel temporal-aware framework tailored for RRG tasks. Libra employs a pre-trained visual transformer encoder, RAD-DINO (Pérez-García et al., 2024), to generate robust image features, which are then refined using a new projector crafted for the temporal awareness, before being fed into the medical large language model (LLM), Meditron (Chen et al., 2023c). Through a two-stage training strategy, Libra aligns temporal visual features with the text embedding space, improving temporal coherence in RRG.

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

[2]E.g., severity and temporal progression of findings.

Our modular approach integrates state-of-the-art open-source pre-trained models for medical image and text processing while introducing a dedicated temporal-aware adapter to align visual and textual modalities within the embedding space. This paper makes the following contributions:

- **Libra**, a temporal-aware MLLM designed to model temporal references and mitigate temporal hallucinations in RRG tasks.
- **Temporal Alignment Connector (TAC)**, comprising the Layerwise Feature Extractor (LFE) and Temporal Fusion Module (TFM), which extracts high-granularity image features from multiple encoder layers and integrates temporal references from the prior study when available.
- **Extensive evaluation** on the MIMIC-CXR dataset, achieving state-of-the-art results on average among similarly scaled MLLMs, with case analysis illustrating Libra's architectural benefits.

## 2 Libra

### 2.1 Model Architecture

Our Libra model follows the standard architecture of MLLMs, such as LLaVA (Liu et al., 2023), comprising an image encoder, a text decoder and a connector module to map visual features into the text embedding space. Figure 2 shows the overall architecture of Libra. Specifically, we utilise a frozen biomedical image encoder, i.e. 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). The text encoder is deployed by Meditron-7B (Chen et al., 2023c), which builds on Llama-2 and is further pre-trained on specialised medical corpora.

To effectively connect the image encoder and LLM, we design a novel Temporal Alignment Connector (TAC) tailored to capture and integrate temporal information from paired images taken at different time points. Meanwhile, when no prior image is available, we employ a dummy prior image, which is simply a copy of the current image, to mitigate spurious references to nonexistent scans, as shown in Figure 2 (bottom). This design enables Libra to effectively manage temporal data (e.g., stable, improved, worsening) and enhances its ability to generate accurate and coherent radiology reports.

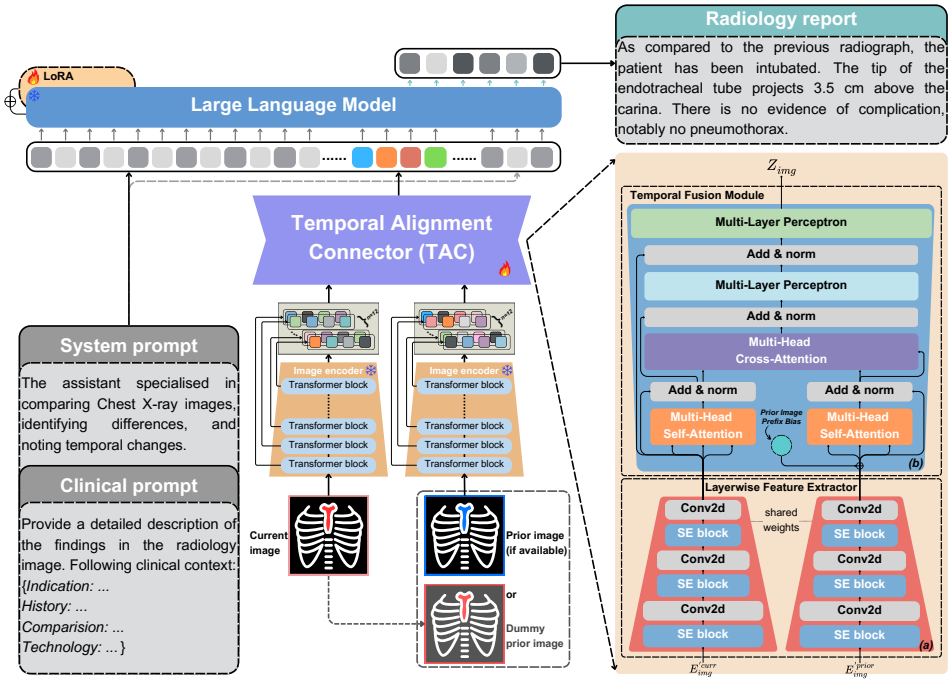


Figure 2: The overall architecture of Libra. The core component, the Temporal Alignment Connector (TAC), processes paired temporal images to enhance temporal reasoning. TAC consists of two key modules: ***(a)*** the Layerwise Feature Extractor (LFE), which aggregates multi-layer image features from the image encoder, and ***(b)*** the Temporal Fusion Module (TFM), which aligns the extracted features and integrates temporal differences before feeding them into the LLM. When no prior image is available, a dummy prior image is used to support temporal modelling, mitigate hallucinations, and prevent spurious references to nonexistent prior studies.

### 2.2 Temporal Alignment Connector

To address the challenges of integrating temporal information and aligning high-granularity visual features for RRG tasks, TAC bridges the image encoder and the LLM. It processes visual features from two temporal snapshots to produce a unified representation sensitive to temporal changes. As shown in Figure 2 (right), TAC includes two key components: the *Layerwise Feature Extractor*, which extracts high-granularity image representations, and the *Temporal Fusion Module*, which integrates temporal references from the prior study.

#### 2.2.1 Layerwise Feature Extractor

To leverage abundant image feature representations encoded by a pre-trained image encoder, we extract image patch token features of all the hidden layers for a given pair of input images. By default, the RAD-DINO image encoder (Pérez-García et al., 2024) has 12 hidden layers and processes $518 \times 518$ input images into $14 \times 14$ patches, generating 1,369 patch token sequences per hidden layer. Rather than relying on a single global feature token (e.g., the $[CLS]$ token), we collect same-dimensional patch embeddings from each layer per image, denoted as $E^{\text{img}} \in \mathbb{R}^{N \times D_{\text{img}}}$, where $N = 1,369$ is the number of patch tokens and $D_{\text{img}}$ is the embedding dimension of the image encoder.

Then, these embeddings are concatenated across all layers as $E'_{\text{img}} = \{E_i^{\text{img}}\}_{i=1}^{n}$, where $n$ is the number of hidden layers. Drawing from the progressive compression strategy in VGG (Simonyan and Zisserman, 2015), our Layerwise Feature Extractor (LFE) reduces dimensionality across layers while preserving critical information. First, 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 and 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 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 features are compressed according to $\{k, j\} \in \{12, 6, 3, 1\}$[3]. Through three stages of progressive compression,

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

we obtain the final patch-level representation:

$$A'_{\text{img}} = Conv2d_6^{12}(SE_6^{12}(E'_{\text{img}})) \quad (1)$$

$$A''_{\text{img}} = Conv2d_3^{6}(SE_3^{6}(A'_{\text{img}})) \quad (2)$$

$$A_{\text{img}} = Conv2d_1^{3}(SE_1^{3}(A''_{\text{img}})) \quad (3)$$

For simplicity, we use $LFE(\cdot)$ to denote the above three stages of compression, which project a given input image $E'_{\text{img}}$ into its feature representation of the fixed dimension, $A_{\text{img}} \in \mathbb{R}^{1\times N\times D_{\text{img}}}$:

$$A_{\text{img}} = LFE(E'_{\text{img}}) \quad (4)$$

By progressively refining each image's representations through multiple stages, the LFE generates a unified and compact feature set suitable for temporal alignment. This design ensures that both high-granularity and global context are retained, as illustrated in ***(a)*** of Figure 2.

### 2.2.2 Temporal Fusion Module

The Temporal Fusion Module (TFM) is inspired by the transformer decoder and is designed to integrate temporal information by leveraging prior images as auxiliary context. It takes as input a paired set of compressed features from both the current and prior images, denoted as $A_{\text{img}}^{\text{curr}}$ and $A_{\text{img}}^{\text{prior}}$, respectively, which are obtained after processing through the LFE. The temporal fusion process is defined as:

$$Z_{\text{img}} = TFM(A_{\text{img}}^{\text{curr}}, A_{\text{img}}^{\text{prior}}) \quad (5)$$

where TFM learns to weigh the current image using prior image features, refining the representation to enhance temporal awareness. The resulting feature sequence, $Z_{\text{img}} \in \mathbb{R}^{N\times d}$, serves as the input to the LLM, where $N$ is the number of patch tokens and $d$ is the hidden dimension of the LLM. This process encapsulates the temporal evolution of the patient's condition, allowing the language model to generate accurate and contextually aware radiology reports.

**Prior Image Prefix Bias** The dataset contains samples with and without a prior image. When a prior image is not available, we set $A_{\text{img}}^{\text{prior}} = A_{\text{img}}^{\text{curr}}$. However, this "dummy prior image" is indistinguishable from a true prior in raw features. To differentiate it, we add a trainable bias, as $b_{\text{prior}}$.

Following the attention scaling techniques for adjusting hidden space degrees of freedom with a chi-square distribution (Vaswani, 2017), a nonlinear scaling function amplifies higher similarity values. The cosine similarity between the current and prior images is scaled with an exponent of $\sqrt[4]{d}$, where $d$ is the hidden dimension of the LLM:

$$b'_{\text{prior}} = b_{\text{prior}} \cdot \left(\frac{cos(A_{\text{img}}^{\text{curr}}, A_{\text{img}}^{\text{prior}}) + 1}{2}\right)^{\sqrt[4]{d}} \quad (6)$$

$$A_{\text{img}}^{\prime\text{prior}} = A_{\text{img}}^{\text{prior}} + b'_{\text{prior}} \quad (7)$$

This nonlinear scaling emphasises high similarity values, modulating the influence of prior image features. When no 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, in Figure 2.

**Transformer Block** The Transformer Block in TFM follows the standard Transformer design but is optimized for handling temporal image pairs. It consists of multi-head self-attention ($SelfAttn$), multi-head cross-attention ($CrossAttn$), and two multi-layer perceptron ($MLP$) sub-layers. As illustrated in ***(b)*** of Figure 2. The paired $(A_{\text{img}}^{\text{curr}}, A_{\text{img}}^{\prime\text{prior}})$ are processed with layer normalization ($LN$) and residual connections:

$$T_{\text{curr}}^{\text{self}} = LN(A_{\text{img}}^{\text{curr}} + SelfAttn(A_{\text{img}}^{\text{curr}}; A_{\text{img}}^{\text{curr}})) \quad (8)$$

$$T_{\text{prior}}^{\text{self}} = LN(A_{\text{img}}^{\prime\text{prior}} + SelfAttn(A_{\text{img}}^{\prime\text{prior}}; A_{\text{img}}^{\prime\text{prior}})) \quad (9)$$

$$T_{\text{img}}^{\text{cross}} = LN(T_{\text{curr}}^{\text{self}} + CrossAttn(T_{\text{curr}}^{\text{self}}; T_{\text{prior}}^{\text{self}})) \quad (10)$$

$$T_{\text{img}}^{\text{out}} = LN(A_{\text{img}}^{\text{curr}} + MLP_{\text{attn}}(T_{\text{img}}^{\text{cross}})) \quad (11)$$

$$Z_{\text{img}} = MLP_{\text{final}}(T_{\text{img}}^{\text{out}}) \quad (12)$$

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

## 2.3 Prompt Design

To enhance Libra's ability to perceive temporal changes and integrate medical information in RRG, we design a structured prompting strategy, consisting of a system prompt and a clinical prompt, as shown in Figure 2 (left). The system prompt enables the LLM to recognise temporal variations, while standard report sections (*Indication History*, *Comparison*, *Technique*) are integrated into the clinical prompt (see Appx. C for a detailed example).

The full prompt is: "Provide a detailed description of the findings in the radiology image. Following clinical context:{...}." There are datasets, e.g. MIMIC-CXR (Johnson et al., 2019b), where the report sections are unavailable. For these datasets, we set the prompt as follows: "Provide a detailed description of the findings in the radiology image." After tokenising and embedding prompts, the refined image features ($Z_{\text{img}}$) are inserted between the system prompt and clinical prompts.

### 2.4 Temporal-aware Training

Libra focuses on frontal-view images, either posterior-anterior (PA) or anterior-posterior (AP), and targets the *Findings* sections of RRG, as these contain the most direct clinical observations. It employs a two-stage training strategy, inspired by recent MLLM fine-tuning techniques (McKinzie et al., 2024), to progressively learn visual feature alignment and temporal information extraction.

**Temporal Feature Alignment** In the first stage, the visual encoder and LLM weights are frozen, while the TAC is trained. This stage focuses on *Findings* and *Impression* generation from paired images and performing CXR-related visual question answering (VQA) tasks to extract high-quality image representations and capture temporal changes.

**Downstream Task Fine-tuning** In the second stage, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2021) to fine-tune the LLM on the *Findings* generation task, while keeping the visual encoder and TAC weights frozen. LoRA achieves performance comparable to full fine-tuning at a substantially lower computational cost. The detailed training configuration, including learning rate schedules and hyperparameters, is provided in Appx. D.

## 3 Experiments

### 3.1 Task and Dataset

**Task Description** We focus on generating the *Findings* section of radiology reports for frontal CXRs, ensuring a fair comparison with prior work. The *Findings* section provides radiologists' observations, encompassing both normal and abnormal findings. While additional sections like *Indication* and *Technique* primarily serve as routine records (e.g., clinical history or specific physician requests), they also assist the model in understanding temporal changes across images. Hence, 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. Although lateral views are occasionally used to supplement anatomical assessments (Islam et al., 2023), they are excluded in this study to maintain consistency with previous research on RRG tasks, such as Chaves et al. (2024) and Hyland et al. (2024). Both current and prior images in our experiments exclusively utilise single frontal views.

**Dataset Description** Libra is trained and evaluated using the MIMIC-CXR dataset (Johnson et al., 2019b) and its derivative datasets, including Medical-Diff-VQA (Hu et al., 2023) and MIMIC-Ext-*MIMIC-CXR-VQA* (Bae et al., 2023). All datasets are split according to the official labels to prevent data leakage. Detailed dataset descriptions and preprocessing steps are in Appx. E.

Following the dataset scaling law utilised in multi-stage MLLM fine-tuning methods (Zhu et al., 2023), we adopt a two-stage training strategy, as noted in Sec. 2.4. The first stage trains TAC on $\sim$1.2M CXR-image text pairs from MIMIC-CXR and its derivatives, including *Findings*, *Impression*, and VQA tasks, enabling it to learn CXR token distributions and image-text relationships. The second stage fine-tunes the model on downstream tasks, refining the LLM to align high-granularity CXR features with the *Findings* section of reports.

Beyond *Findings* section generation, the first stage incorporates *Impression* section and VQA tasks. The *Impression* section, which summarises diagnoses and proposes further investigations (Babar et al., 2021), facilitates alignment between CXRs and their textual descriptions. We use the same system and clinical prompts as for *Findings*, replacing 'Findings' with 'Impression'. For VQA, the system prompts remain unchanged, while clinical prompts are adapted to address medical-specific questions, guiding caption generation. These VQA tasks refine the MLLM's biomedical vocabulary usage and strengthen image-text alignment.

### 3.2 Evaluation Metrics

We evaluate the generated reports using lexical and radiology-specific metrics, adhering to established protocols. Lexical metrics include ROUGE-L (Lin, 2004), BLEU-$\{1, 4\}$ (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) and BERT (Devlin et al., 2019). Radiology-specific metrics include RadGraph-F1 (Jain et al., 2021), $\text{RG}_{\text{ER}}$ (Delbrouck et al., 2022a), F1-CheXpert (Irvin et al., 2019), CheXbert vector similarity (Smit et al., 2020a), and RadCliQ (Yu et al., 2022) version 0.

These clinical metrics typically emphasise the accuracy of medical findings, prioritising the detection of clinically relevant entities. However, they do not evaluate the model's ability to capture temporal information. Therefore, we introduce the temporal entity F1 score ($F1_{\text{temp}}$) to assess this aspect. In particular, the temporal entity F1 score specifically measures the accuracy of entities related to progression over time described in the report [4].

[4] Full metric descriptions, including $F1_{\text{temp}}$, are in Appx. F.

| 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 | **36.2** (+18.3%) |
| BLEU-1 | 35.4 | 16.9 | 16.4 | 32.3 | 38.1 | 39.2 | **51.2** (+30.6%) |
| BLEU-4 | 14.9 | 4.7 | 17.8 | 11.5 | 15.4 | 14.2 | **24.3** (+36.5%) |
| METEOR | 35.3 | – | – | – | – | 33.3 | **48.7** (+38.0%) |
| **Clinical:** | | | | | | | |
| RadGraph-F1 | 19.1 | – | – | 26.7 | – | 24.3 | **32.4** (+21.3%) |
| $RG_{ER}$ | 23.8 | 20.5 | 13.2 | – | 29.4 | 29.6 | **36.9** (+25.0%) |
| $RadCliQ_0(\downarrow)$ | 3.30 | – | – | – | – | 3.10 | **2.76** (+11.0%) |
| CheXbert vector | 36.9 | – | – | – | – | 44.0 | **46.3** (+5.2%) |
| *CheXpert-F1:* | | | | | | | |
| Micro-F1-14 | 42.7 | 39.3 | 35.5 | 53.6 | **57.3** | 55.7 | 55.3 (-3.4%) |
| Macro-F1-14 | 26.9 | 24.7 | 20.4 | 39.8 | 39.5 | 38.6 | **40.2** (+1.1%) |
| Micro-F1-5 | 43.9 | 41.2 | 25.8 | 57.9 | 57.4 | 56.0 | **58.9** (+1.8%) |
| Macro-F1-5 | 36.3 | 34.5 | 19.6 | 51.6 | 47.7 | 47.7 | **52.6** (+2.0%) |

Table 1: Findings generation performance on the MIMIC-CXR test split. ‡ denotes results from Chaves et al. (2024), while '–' indicates missing data. The best performances in **bold**, and the second-best scores are underlined. Metrics where lower values are better are marked with '↓'. Percentage (%) shows improvement over the best existing model.

**Temporal Entity F1** Building on the work of Bannur et al. (2023), we set a reward list comprising common radiology-related keywords indicative of temporal changes. Temporal entities are then extracted from both the ground truth $(E_{\text{gt}})$ and the generated reports $(E_{\text{gr}})$ without applying stemming or lemmatization, preserving the precision of temporal descriptions. After extraction, we compute precision $(P_{\text{temp}})$ and recall $(R_{\text{temp}})$, which are subsequently used to calculate the $F1_{\text{temp}}$, defined as the harmonic mean of precision and recall (Van Rijsbergen, 1974), also known as the $F1$ score.

$$F1_{\text{temp}} = (1+\beta^2) \cdot \frac{P_{\text{temp}} R_{\text{temp}}}{\beta^2 \cdot P_{\text{temp}} + R_{\text{temp}}} \quad (13)$$

$$P_{\text{temp}} = \frac{|E_{\text{gr}} \cap E_{\text{gt}}| + \epsilon}{|E_{\text{gr}}| + \epsilon} \quad (14)$$

$$R_{\text{temp}} = \frac{|E_{\text{gr}} \cap E_{\text{gt}}| + \epsilon}{|E_{\text{gt}}| + \epsilon} \quad (15)$$

where $\epsilon$ is a small value, set to a default of $1 \times 10^{-10}$, 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).

### 3.3 Baselines

While the MIMIC-CXR dataset provides an "official" test split, strict comparisons with prior studies are challenging due to differences in 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., 2021). Such variations in test set distributions can significantly impact the reported results (Park et al., 2024). To ensure fairness [5], 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).

Recent concurrent work, such as M4CXR (Park et al., 2024), employs multi-turn chain-of-thought prompting (Wei et al., 2023) for report generation, which differs from our task setup. Additionally, we do not compare with MAIRA-2 (Bannur et al., 2024), a model designed for grounded radiology report generation incorporating lateral views and prior study reports for each subject within the input prompt. Bannur et al. (2024) emphasises a positive transfer between this distinct task setup and standard RRG, which falls beyond our study's scope. For comparison and discussion of the latest concurrent and non-LLM-based models, see Appx. G.

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

### 3.4 Results

From Table 1, Libra[6] achieves competitive results across most traditional lexical and clinical metrics, excelling in ROUGE-L, BLEU, METEOR, and RadGraph-based scores. It also leads in the radiologist-aligned RadCliQ metric and CheXbert vector similarity. In the CheXpert classification, it attains the highest Macro-F1 scores and a competitive Micro-F1. Overall, Libra demonstrates robust performance in RRG by effectively leveraging temporal information, with only minor gaps in select clinical metrics. These results highlight the effectiveness of its TAC in capturing temporal contexts and generating clinically relevant radiology reports.

[5] The test set includes 2,461 frontal-view samples.

[6] Libra was tested on single-image inputs without priors for fair comparison with models lacking temporal modelling.

| 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 | 22.20 (-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 (-0.24%) | 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%) |

Table 2: 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*.

## 4 Ablation Studies

We conducted ablation studies on Libra's key components, evaluating module and dataset configurations. All experiments were performed on the MIMIC-CXR test split for the *Findings* generation, with prior images included by default and consistent hyperparameters during training and inference.

**Does the Temporal Alignment Connector improve model performance?** To evaluate the impact of TAC on Libra's performance in RRG, we used a model initialised with the RAD-DINO (Pérez-García et al., 2024) image encoder, TAC, and Meditron-7b (Chen et al., 2023c) as the LLM. The baseline (*Libra-1*) was conducted by fine-tuning only the TAC for the *Findings* generation task. As shown in Table 2, we performed ablation studies by progressively removing different TAC components, including the Temporal Fusion Module (TFM), Layerwise Feature Extractor (LFE), Prior Image Prefix Bias (PIPB), and the entire TAC.

Removing TFM restricted the model to single-image processing, akin to LLaVA (Liu et al., 2023), but with a four-layer MLP for aligning image features with the LLM's hidden dimensions. Without LFE, the model used the penultimate layer of the encoder. Removing PIPB excluded the mechanism for differentiating true and dummy prior images. Finally, removing the entire TAC left the model reliant solely on the image encoder and LLM.

The results indicate that removing any TAC submodule leads to performance declines across all metrics compared to *Libra-1*. TFM removal caused a notable drop in the $F1_{temp}$ score (↓>2%), highlighting its role in capturing temporal information. LFE removal significantly decreased RadGraph-related scores, underscoring its importance in extracting detailed image features. PIPB removal impacted clinical metrics more than lexical metrics, indicating its role in enhancing clinical relevance. Complete TAC removal led to substantial declines in all metrics, demonstrating its critical role in integrating image details and temporal information. The evaluation confirms that TAC plays a vital role in improving Libra's ability to generate high-quality, temporally aware radiology reports.

For additional ablation studies exploring TAC's contributions, including its impact under general-domain and radiology-specific pre-trained models, its performance after the second training stage, and its robustness through extended fine-tuning and diverse conditions, and an analysis of whether incorporating temporal information improves Libra's performance in RRG tasks, please refer to Appx. H.

**Are additional *Impression* and VQA datasets necessary during the feature alignment?** To assess the impact of incorporating additional datasets during the first stage of training, we compared a model (*Libra-f*) trained solely on the *Findings* data with Libra, which also used *Impression* and VQA data for feature alignment, as shown in Table 3.

After the first stage, Libra outperformed *Libra-f* in lexical metrics but showed a slight decline in clinical scores. This decline stems from VQA tasks emphasizing fine-grained, grounded descriptions rather than holistic findings. VQA focuses on individual symptoms, whereas *Findings* integrates multiple normal and abnormal observations, affecting $F1_{temp}$ by reducing identified temporal entities.

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

Table 3: Ablation results for dataset configurations. $^{\triangle}$ denotes improvement, while $^{\blacktriangledown}$ indicates decline.

In the second stage, fine-tuning on *Findings* restored balance, further improving performance. These results indicate that additional datasets enhance Libra's RRG ability, while second-stage fine-tuning ensures well-rounded report generation.

## 5 Performance Analysis

We qualitatively assess Libra's ability to generate temporally consistent radiology reports.

**Cases without Prior Image** As shown in Figure 3 ***(a)***, Libra produced detailed descriptions beyond the ground truth, identifying "sternal wires" and their type. This demonstrates its capability to deliver clinically relevant information without spurious referencing nonexistent prior studies.

**Cases with Prior Image** In Figure 3 ***(b)***, new abnormalities such as pleural effusion and pneumonia appeared in the current image. Without a prior image, Libra correctly described the present findings but did not infer disease progression, avoiding spurious references while still suggesting further investigations. When the prior image was considered, Libra effectively captured these progressive changes, provided detailed descriptions, and explicitly referenced the comparison. This facilitated a clear understanding of temporal changes and more accurate descriptions of disease progression.

**Evaluating Temporal Consistency** To assess temporal reasoning, we swapped image order, using the prior image as the current image and vice versa. The generated report then reflected an improved patient condition, aligning with the reversed input sequence but contradicting the ground truth of the original current image. Notably, the report

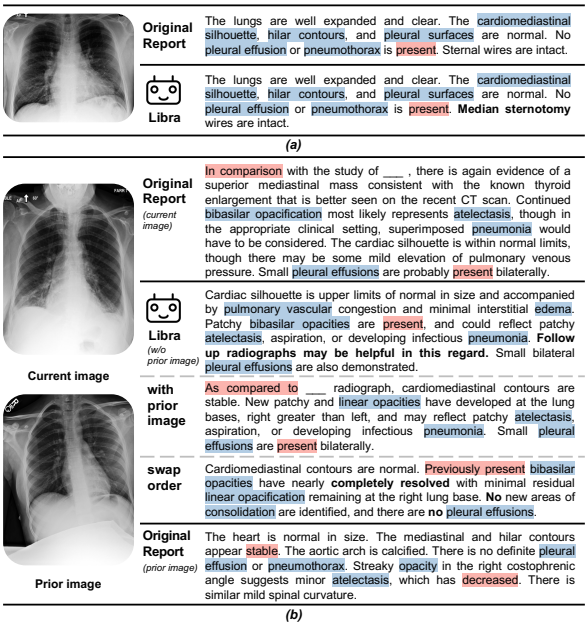


Figure 3: Radiological symptoms, while temporal changes are in red. Key highlights presented in **bold**. Heatmap analysis is available in Appx. I.

closely resembled the original description of the prior image, as shown at the bottom of Figure 3 ***(b)***. This indicates that Libra can effectively adapt to both temporal contexts, generating accurate and contextually consistent reports that simulate the conditions of standard clinical practice.

## 6 Conclusion

In this study, we introduced Libra, a temporal-aware multimodal large language model tailored for chest X-ray report generation tasks. Libra employs a two-stage training framework, leveraging a radiology-specific image encoder and language model connected via the Temporal Alignment Connector, enabling seamless integration of visual and textual modalities. Trained solely on the open-access MIMIC-CXR dataset (Johnson et al., 2019b), Libra demonstrates notable performance gains across key metrics compared to similarly scaled models. Through qualitative and quantitative analysis, we showed that Libra effectively utilises temporal relationships between current and prior scans, addressing challenges such as hallucinations in referencing prior studies. This highlights Libra's ability to generate clinically accurate and temporally consistent radiology reports, setting a new paradigm for multimodal medical AI research.

Future work will focus on expanding Libra's clinical applicability by incorporating diverse imaging modalities and enhancing temporal reasoning capabilities, and extending it in an agentic way.

## Limitations

Despite Libra's ability to model temporal paired images for radiology report generation (RRG), certain limitations remain. First, Libra relies on single prior images for temporal modelling, whereas clinical practice often involves multiple prior scans with varied intervals and angles. Extending the model to handle multiple temporally sequenced images remains an open challenge. Second, our study is based on a single-source dataset with inherent biases in patient demographics and imaging protocols, which may limit generalizability across broader clinical settings. Lastly, while Libra is designed for CXR-based RRG, its applicability to other imaging modalities (e.g., CT, MRI) and integration with structured medical knowledge remains unexplored. For a detailed discussion of these limitations and future directions, see Appx. J.

## Ethics Statement

This work presents Libra, a model designed to enhance radiology report generation by integrating temporal and visual information. While Libra has the potential to improve clinical workflows, reduce radiologist workload, and enhance diagnostic consistency, its deployment must be approached with caution to ensure ethical and responsible use.

Our research exclusively utilises the publicly available and "de-identified" MIMIC-CXR dataset (Johnson et al., 2019b), in accordance with its official documentation, ensuring adherence to ethical and privacy standards under CITI Data or Specimens Only Research certification. By relying solely on open datasets, we prioritise transparency and reproducibility, aligning with best practices in ethical AI research.

This work is intended to support, not replace, medical professionals, ensuring it serves as a complementary tool within clinical practice. While the societal implications are largely positive, further validation across diverse patient populations and healthcare systems is necessary to address potential biases inherent in the dataset. Additionally, it is crucial to mitigate the risks of over-reliance on AI systems, which could inadvertently undermine human oversight or exacerbate healthcare disparities.

Future efforts will aim to extend the model's capabilities to encompass multiple imaging modalities and broader datasets, ensuring greater generalisability, fairness, and adaptability across diverse clinical settings.

## References


Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering . In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, Los Alamitos, CA, USA. IEEE Computer Society.

Zaheer Babar, Twan van Laarhoven, Fabio Massimo Zanzotto, and Elena Marchiori. 2021. Evaluating diagnostic content of ai-generated radiology reports of chest x-rays. *Artificial Intelligence in Medicine*, 116:102075.

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. 2023. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. *Preprint*, arXiv:2310.18652.

Oishi Banerjee, Hong-Yu Zhou, Subathra Adithan, Stephen Kwak, Kay Wu, and Pranav Rajpurkar. 2024. Direct preference optimization for suppressing hallucinated prior exams in radiology report generation. *Preprint*, arXiv:2406.06496.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

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. 2024. Maira-2: Grounded radiology report generation. *Preprint*, arXiv:2406.04449.

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. 2023. Learning to exploit temporal structure for biomedical vision-language processing. *Preprint*, arXiv:2301.04558.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797.

Yiming Cao, Lizhen Cui, Lei Zhang, Fuqiang Yu, Zhen Li, and Yonghui Xu. 2023. Mmtn: Multi-modal memory transformer network for image-report consistent medical report generation. *Proceedings*

*of the AAAI Conference on Artificial Intelligence*, 37(1):277–285.

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, and 8 others. 2024. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. *Preprint*, arXiv:2403.08002.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023a. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56.

Qiuhui Chen, Xinyue Hu, Zirui Wang, and Yi Hong. 2023b. Medblip: Bootstrapping language-image pre-training from 3d medical images and texts. *Preprint*, arXiv:2305.10799.

Weixing Chen, Yang Liu, Ce Wang, Jiarui Zhu, Shen Zhao, Guanbin Li, Cheng-Lin Liu, and Liang Lin. 2024a. Cross-modal causal intervention for medical report generation. *Preprint*, arXiv:2303.09117.

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. 2023c. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *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)*, pages 5904–5914, Online. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2022. Generating radiology reports via memory-driven transformer. *Preprint*, arXiv: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. 2024b. Chexagent: Towards a foundation model for chest x-ray interpretation. *Preprint*, arXiv: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. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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, and 48 others. 2022. Palm: Scaling language modeling with pathways. *Preprint*, arXiv:2204.02311.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis P. Langlotz. 2022b. Improving the factual correctness of radiology report generation with semantic rewards. *Preprint*, arXiv: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. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Xinpeng Ding, Yongqiang Chu, Renjie Pi, Hualiang Wang, and Xiaomeng Li. 2024. HiA: Towards Chinese Multimodal LLMs for Comparative High-Resolution Joint Diagnosis . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15012. Springer Nature Switzerland.

Emil Fischer. 1894. Einfluss der configuration auf die wirkung der enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993.

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. 2018. Structured reporting in radiology. *Academic radiology*, 25(1):66–73.

Tiancheng Gu, Dongnan Liu, Zhiyuan Li, and Weidong Cai. 2023. Complex organ mask guided radiology report generation. *Preprint*, arXiv:2311.02329.

Dan Hendrycks and Kevin Gimpel. 2023. Gaussian error linear units (gelus). *Preprint*, arXiv:1606.08415.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. 2023a. RECAP: Towards precise radiology report generation via dynamic disease progression reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2134–2147, Singapore. Association for Computational Linguistics.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023b. Organ: Observation-guided radiology report generation via tree reasoning. *Preprint*, arXiv:2306.06466.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Jie Hu, Li Shen, Samuel Albanie, Gang Sun, and Enhua Wu. 2019. Squeeze-and-excitation networks. *Preprint*, arXiv:1709.01507.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, Liangchen Liu, Kazuma Kobayashi, Tatsuya Harada, Ronald M. Summers, and Yingying Zhu. 2023. 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, page 4156–4165, New York, NY, USA. Association for Computing Machinery.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. *Preprint*, arXiv: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. 2024. Maira-1: A specialised large multimodal model for radiology report generation. *Preprint*, arXiv: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. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Preprint*, arXiv:1901.07031.

S. K. M Shadekul Islam, MD Abdullah Al Nasim, Ismail Hossain, Md Azim Ullah, Kishor Datta Gupta, and Md Monjur Hossain Bhuiyan. 2023. Introduction of medical imaging modalities. *Preprint*, arXiv: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. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. *Preprint*, arXiv:2106.14463.

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

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

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. 2019a. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *Preprint*, arXiv:1901.07042.

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

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

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

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023b. Dynamic graph enhanced contrastive learning for chest x-ray report generation. *Preprint*, arXiv:2303.10323.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Fenglin Liu, Shen Ge, and Xian Wu. 2021a. Competence-based multimodal curriculum learning for medical report generation. In *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)*, pages 3001–3012, Online. Association for Computational Linguistics.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021b. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation . In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13748–13757, Los Alamitos, CA, USA. IEEE Computer Society.

Fenglin Liu, Changchang Yin, Xian Wu, Shen Ge, Ping Zhang, and Xu Sun. 2021c. Contrastive attention for automatic chest X-ray report generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 269–280, Online. Association for Computational Linguistics.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. *Preprint*, arXiv:1904.02633.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250.

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, and 13 others. 2024. Mm1: Methods, analysis and insights from multimodal llm pre-training. *Preprint*, arXiv:2403.09611.

Xin Mei, Rui Mao, Xiaoyan Cai, Libin Yang, and Erik Cambria. 2024. Medical report generation via multimodal spatio-temporal fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 4699–4708, New York, NY, USA. Association for Computing Machinery.

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

Yasuhide Miura, Yuhao Zhang, Emily Bao Tsai, Curtis P. Langlotz, and Dan Jurafsky. 2021b. Improving factual completeness and consistency of image-to-text radiology report generation. *Preprint*, arXiv:2010.10042.

Reabal Najjar. 2023. Redefining radiology: a review of artificial intelligence integration in medical imaging. *Diagnostics*, 13(17):2760.

Aaron Nicolson, Jason Dowling, and Bevan Koopman. 2023. Improving chest X-ray report generation by leveraging warm starting. *Artificial Intelligence in Medicine*, 144:102633.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive transformer-based generation of radiology reports. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2824–2832, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv: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, and 7 others. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

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

Jonggwon Park, Soobum Kim, Byungmu Yoon, Jihun Hyun, and Kyoyun Choi. 2024. M4cxr: Exploring multi-task potentials of multi-modal large language models for chest x-ray interpretation. *Preprint*, arXiv:2408.16213.

Linus Pauling, Robert B. Corey, and H. R. Branson. 1951. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205–211.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. 2023. Radialog: A large vision-language model for radiology report generation and conversational assistance. *Preprint*, arXiv:2311.18681.

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. 2024. Raddino: Exploring scalable medical image encoders beyond text supervision. *Preprint*, arXiv:2401.10815.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 448–458, Dublin, Ireland. Association for Computational Linguistics.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-Critical Sequence Training for Image Captioning . In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195, Los Alamitos, CA, USA. IEEE Computer Society.

Santosh Sanjeev, Fadillah Adamsyah Maani, Arsen Abzhanov, Vijay Ram Papineni, Ibrahim Almakky, Bartłomiej W. Papież, and Mohammad Yaqub. 2024. Tibix: Leveraging temporal information for bidirectional x-ray and report generation. *Preprint*, arXiv:2403.13343.

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

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *Preprint*, arXiv:1409.1556.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. 2024. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, page 1–20.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020a. Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, Online. Association for Computational Linguistics.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020b. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Xiao Song, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Pengxu Wei. 2022. Cross-modal contrastive attention model for medical report generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2388–2397, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

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, and 13 others. 2023. Towards generalist biomedical ai. *Preprint*, arXiv:2307.14334.

Cornelis Joost Van Rijsbergen. 1974. Foundation of evaluation. *Journal of Documentation*, Volume 30(Issue 4):365–373.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Jun Wang, Abhir Bhalerao, and Yulan He. 2022. Cross-modal prototype driven network for radiology report generation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 563–579. Springer.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M. Summers. 2018. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. *Preprint*, arXiv:1801.04334.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. *Preprint*, arXiv:2304.02211.

James D Watson and Francis HC Crick. 1953. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

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. 2021. Chest imagenome dataset for clinical reasoning. *Preprint*, arXiv:2108.00316.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. 2024. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, 117:102888.

Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao. 2022a. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80:102510.

Shuxin Yang, Xian Wu, Shen Ge, S. Kevin Zhou, and Li Xiao. 2022b. Radiology report generation with a learned knowledge base and multi-modal alignment. *Preprint*, arXiv:2112.15011.

Di You, Fenglin Liu, Shen Ge, Xiaoxia Xie, Jing Zhang, and Xian Wu. 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III*, page 72–82, Berlin, Heidelberg. Springer-Verlag.

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. 2022. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*.

Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, and Jian Wu. 2024. Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. *Preprint*, arXiv:2403.04290.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024a. Vision-language models for vision tasks: A survey. *Preprint*, arXiv:2304.00685.

Ke Zhang, Hanliang Jiang, Jian Zhang, Qingming Huang, Jianping Fan, Jun Yu, and Weidong Han. 2024b. Semi-supervised medical report generation via graph-guided hybrid feature consistency. *Trans. Multi.*, 26:904–915.

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, and 2 others. 2024c. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *Preprint*, arXiv:2303.00915.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020a. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. 2020b. When radiology report generation meets knowledge graph. *Preprint*, arXiv:2002.08277.

Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J. Topol, and Pranav Rajpurkar. 2024. A generalist learner for multifaceted medical image interpretation. *Preprint*, arXiv:2405.07988.

Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. 2023. Advancing radiograph representation learning with masked record modeling. *Preprint*, arXiv:2301.13155.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *Preprint*, arXiv:2304.10592.

## Appendix Contents

# A Related Work

## A.1 Radiology Report Generation

Radiology report generation (RRG) aims to address the long-tail distribution of observations in chest X-rays (CXRs) and produce fine-grained descriptions of clinical findings, making it a key objective in automated medical imaging analysis (Wang et al., 2018).

Early RRG systems relied on recurrent neural networks (RNNs) (Liu et al., 2019), which have since been largely replaced by transformer-based 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). These models excel at language generation, offering substantial improvements in fluency and factual accuracy.

To further enhance clinical accuracy, some methods incorporate reinforcement learning (RL) to optimise for task-specific rewards, such as capturing "clinically relevant" features (Liu et al., 2019; Irvin et al., 2019) or maintaining logical consistency (Miura et al., 2021a; Delbrouck et al., 2022a). However, these approaches often rely on external tools like CheXbert (Smit et al., 2020a) or RadGraph (Jain et al., 2021), adding complexity to the optimisation process.

Recent advancements in LLMs have shown that plain auto-regressive language modelling can achieve strong performance in RRG tasks. However, RL-based objectives and task-specific optimisations remain complementary, offering additional opportunities for improvement. Research on leveraging temporal information in RRG tasks can be broadly categorised into LLM-based and non-LLM-based methods, each presenting distinct advantages and challenges.

## A.2 LLM-based Model

LLM-based models have achieved significant success in the RRG task, primarily due to advancements in 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 connects the encoder's outputs to the LLM. Most existing adapters use single-layer hidden representations (e.g., the last or penultimate layer) from pre-trained image encoders, limiting their ability to integrate features from multiple images effectively.

In end-to-end training, LLM-based models handle multiple image inputs by concatenating them with textual prompts, forming a composite input to the LLM. For instance, the input format is often structured as "<Current Image Placeholder> + <Prior Image Placeholder> + <Prompt>". However, this approach provides limited guidance on the relationship between the images within the prompt. Ding et al. (2024) proposed the High-Resolution Instruction-Aware Adapter (HiA) to refine image-text representations, improving the model's ability to follow textual prompts with multiple images. While this enhances instruction adherence, it does not explicitly model relationships between paired images.

In contrast to this vanilla approach, Libra explicitly models temporal relationships in paired images through its Temporal Alignment Connector (TAC). Instead of simply concatenating images in the LLM's latent space, TAC leverages all hidden-layer features from the image encoder to provide richer feature representations. By directly modelling temporal dynamics, Libra enables more precise and context-aware radiology report generation.

## A.3 non-LLM-based Model

Non-LLM-based models typically employ transformer encoder-decoder architectures or their variants, which often require separate training for individual modules. These approaches handle "single-" and "double-" image inputs by symbolically differentiating tasks and employing distinct architectures tailored for each input type. Additionally, they frequently incorporate extra information such as prior reports, symptom labels, and knowledge graphs.

For instance, Serra et al. (2023) uses symbolic alignment in its Longitudinal Projection Module along with a separately trained BERT-based (Devlin et al., 2019) text generator. RECAP (Hou et al., 2023a) implements a two-stage training process: classification tasks followed by report generation, leveraging a transformer encoder-decoder with symbolic task differentiation. TiBiX (Sanjeev et al., 2024) incorporates causal attention layers and learnable padding tokens to handle cases without prior images, while 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 gen-

erally outperform other types of models (i.e. non-LLM-based) in the RRG task, papers on non-LLM-based models or those using small language models (SLMs) typically do not compare their methods with LLM-based approaches. Nonetheless, we conducted comparisons and discussions to reaffirm this observation, as detailed in Appx. G.2.

### A.4 Radiological Image Representation

Radiology-specific pre-trained image encoder models are essential for RRG tasks due to the unique characteristics of radiological images, which fall outside the distribution of general-domain image models (Pérez-García et al., 2024).

Several notable advancements have been made in this domain. Zhou et al. (2023) proposed Masked Record Modeling (MRM), a unified framework combining self-supervision with radiology report supervision to enhance radiograph representation learning. Similarly, BioViL-T (Bannur et al., 2023) employs a CNN-Transformer hybrid architecture to model multimodal relationships and leverage temporal structures for tasks such as disease progression classification and report generation. In addition, BiomedCLIP (Zhang et al., 2024c) is a multimodal biomedical foundational model pretrained across diverse 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. Recent works have extensively applied RAD-DINO to RRG tasks, including MAIRA-2 (Bannur et al., 2024) and M4CXR (Park et al., 2024). Notably, Pérez-García et al. (2024) demonstrated that RAD-DINO outperforms other image encoders in RRG tasks.

Building on this evidence, our model incorporates RAD-DINO as its image encoder to ensure high-quality radiological image representations, providing a robust foundation for downstream RRG tasks.

## B Research Objectives

### B.1 Temporal Information

Temporal changes are critical for understanding disease progression. In radiology, paired images and their corresponding reports document subtle evolutions of symptoms over time. This temporal information is often captured by comparing current scans with prior ones to highlight symptom evolution or newly identified findings.

The relative positioning of scans within the timeline determines the extent of temporal information. Therefore, the relative timing between scans is key: when the prior scan is recent, reported changes tend to be minimal; conversely, an older prior scan reveals more pronounced differences.

Importantly, while temporal context enriches the diagnostic narrative, it does not alter the factual observations present in the current scan—it merely provides additional layers of interpretative insight.

### B.2 Research Aims

This study aims to enhance radiology report generation (RRG) by effectively incorporating temporal information into the modelling process. In clinical practice, chest X-ray (CXR) analysis often depends on comparing the current scan with the prior image to capture disease progression and evolution. Our primary objective is to leverage these temporal cues to generate more accurate, context-aware radiological reports that faithfully reflect both stable conditions and clinically significant changes.

Unlike previous LLM-based models (discussed in Appx. A.2), which depend on the LLM to infer temporal information solely from text, our approach explicitly models temporal relationships at the architectural level. Inspired by the principle **"structure determines function"** (Fischer, 1894; Pauling et al., 1951; Watson and Crick, 1953), we introduce the Temporal Alignment Connector (TAC), a dedicated module designed to capture temporal dynamics. Details are provided in Sec. 2.2.

### B.3 Research Scope

This study focuses on frontal chest X-rays, treating each examination per image while incorporating a single prior image as an auxiliary input when available. Rather than modelling patient-level longitudinal history, our goal is to generate a report for the current image while leveraging temporal information from one preceding scan. To ensure fairness in benchmarking, Libra was evaluated on single-image inputs without priors (see Table 1). Yet, temporal information remains implicitly present through several factors:

- Explicit temporal states (e.g., "stable" or "unstable") are frequently described in reports.
- Latent temporal progression exists in datasets, as prior studies influence diagnostic phrasing.

• The absence of a prior image itself constitutes a temporal scenario, representing an extreme case where the patient's condition is assumed stable due to a lack of comparative reference.

Our model can effectively 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"; our model should then generate a report reflecting this stability. When no prior image is available, we employ a dummy prior image (a copy of the current image) to maintain input consistency and mitigate spurious references to nonexistent priors.

However, in clinical practice, patients often undergo multiple prior scans, sometimes from different orientations, providing a more complex temporal context. This lies beyond the scope of our current study, and a detailed discussion of such scenarios is provided in Appx. J.

## C Prompt Example

We selected examples from the MIMIC-CXR (Johnson et al., 2019b) dataset and synthesised them using GPT-4 (OpenAI et al., 2024) to ensure ethical compliance, as illustrated in Table 4. Following the rule-based approach by Hyland et al. (2024), we extracted key sections from the report of the current image. Each example combines a fixed system prompt with a dynamic clinical prompt tailored to the current scan. We utilised four clinical instructions from the original report: {*Indication*}, {*History*}, {*Comparison*}, and {*Technique*}. In contrast, MAIRA-2 (Bannur et al., 2024), which incorporates prior image reports, our approach focuses exclusively on the current image's context, maintaining a clear distinction from prior study information of the report.

## D Training Configuration

Libra is trained using a standard auto-regressive language modelling loss (cross-entropy). For this study, we employ Meditron-7b (Chen et al., 2023c) 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 using DeepSpeed optimization (Rajbhandari et al., 2020) with ZeRO-2 for stage 1 and ZeRO-3 for stage 2, and BF16 precision is 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 (∼385 hours) with a learning rate of $2\times10^{-5}$. In the second stage, the model is trained for 3 epochs (∼213 hours) 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.

**Note:** Prior work, especially in the medical domain, typically employs full model fine-tuning for RRG tasks. However, due to hardware constraints, we can only adopt a lightweight training technique for parameter-efficient adaptation. As a result, our approach may underperform full model fine-tuning strategies in the second stage, despite maintaining computational efficiency.

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

Table 4: Examples of Libra's system and clinical prompts for *Findings* section generation in RRG task.

| Dataset | Task Type | # Samples | | | % Has Prior | | |
|---|---|---|---|---|---|---|---|
| | | Train (%) | Valid (%) | Test (%) | 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 |

Table 5: Datasets used for training and evaluating Libra include statistics on the proportion of samples that contain prior images. The first stage uses the full dataset, while the second stage fine-tunes for downstream tasks.

## E Datasets Description

**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 (Johnson 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, studies without extractable *Findings* are discarded, while other missing sections are permitted. The same approach is applied to the *Impression* section generation task. In all our experiments, we adhere to the official MIMIC-CXR dataset split.

Meanwhile, 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. As shown in Table 5, 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 extends MIMIC-CXR for VQA tasks tailored to CXRs. It includes questions generated from 48 unique templates covering seven content types: presence, anatomy, attribute, abnormality, size, plane, and gender, as shown in Table 5. 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 (see Appx. B) for the RRG task. After thoroughly evaluating other datasets, we found them **unsuitable** for the following reasons:

**CheXpert** (Irvin et al., 2019) 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., 2021) Although derived from MIMIC-CXR (Johnson et al., 2019b), it does not follow the official split, raising concerns about potential data leakage between training, validation, and test sets.

Meanwhile, the following two datasets were processed using GPT-4 (OpenAI et al., 2024) to eliminate hallucinated references to prior exams. While this prevents erroneous comparisons, it also removes essential temporal information originally present in the reports, potentially affecting tasks that rely on temporal reasoning.

**LLaVA-Rad MIMIC-CXR Dataset** (Chaves et al., 2024) This dataset was refined using GPT-4 (OpenAI et al., 2024) through a structured text-cleaning pipeline. The process involved: (1) correcting typographical errors and split words, (2) removing redundant or repeated phrases to improve clarity, (3) eliminating explicit temporal references (e.g., "Compared to the prior study, no significant interval change was noted") to ensure the report focuses exclusively on the current image, and (4) restructuring content into standardised sections, including *Indication*, *Findings*, and *Impression*.

**ReXPref-Prior Dataset** (Banerjee et al., 2024) A modified version of MIMIC-CXR (Johnson et al., 2019b) in which GPT-4 (OpenAI et al., 2024) systematically removes all references to prior exams from both the *Findings* and *Impression* sections. While this adjustment prevents spurious prior-study references, it also eliminates crucial temporal context, limiting its suitability for applications requiring longitudinal assessment of disease progression.

# F Evaluation Metrics

## F.1 Lexical Metrics

We employed standard 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 and 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.

## F.2 Clinical 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** 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, $\text{RG}_{\text{ER}}$ (Delbrouck et al., 2022b), which matches entities based on their text, type, and whether they have at least one relation[7].

**CheXpert F1** This set of metrics utilizes the CheXbert automatic labeler (Smit et al., 2020a) to extract "present", "absent", or "uncertain" labels for each of the 14 CheXpert pathologies (Irvin et al., 2019) from the generated reports and their

[7] $\text{RG}_{\text{ER}}$ is implemented as `F1RadGraph` with `reward=partial` by the radgraph package.

| 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** |

Table 6: Evaluation of candidate reports using the Temporal Entity F1 score ($F1_{\text{temp}}$). Descriptions of temporal changes are marked.

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 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., 2020a).

**RadCliQ** In addition, we utilise RadCliQ (Radiology Report Clinical Quality) (Yu et al., 2022), 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, $\text{RadCliQ}_0$, and RadGraph-F1 metrics are calculated using the code released by Yu et al. (2022).

### F.3 Temporal Entity F1

We introduced $F1_{\text{temp}}$, a metric specifically designed to detect temporal entities reflecting changes over time. Unlike traditional lexical or radiology-specific metrics, $F1_{\text{temp}}$ evaluates the quality of temporal information in radiology reports.

As shown in Table 6, the differences in lexical (ROUGE-L (Lin, 2004)) and clinical (RadGraph-F1 (Jain et al., 2021)) metrics between the two candidates are relatively smaller compared to the $F1_{\text{temp}}$ score. This demonstrates 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.

## G Analysis of Concurrent Work and Non-LLM-based Models

### G.1 Discussion on Performance with Radiology Foundation Models

As shown in Table 7, these models belong to the category of radiology foundation models.

DaDialog (Pellegrini et al., 2023) is a conversational MLLM designed for a broad range of dialogue-based medical assistance tasks. To enhance structured findings extraction, it employs the publicly available CheXbert model (Smit et al., 2020b) to extract symptom labels from scans, facilitating a structured representation of findings.

MedVersa (Zhou et al., 2024) and M4CXR (Park et al., 2024) support a diverse set of tasks, including medical report generation, visual grounding, and visual question answering. These models aim to provide general-purpose multimodal medical assistance by leveraging vision-language pre-training strategies.

MAIRA-2 (Bannur et al., 2024) specialises in grounded radiology report generation, which differs from traditional report generation tasks by requiring explicit image-level localization of findings and symptoms. Grounded radiology reporting, as defined by Bannur et al. (2024), structures the report as a list of sentences, where each sentence: (1) is linked to zero or more spatial image annotations, and (2) describes at most a single finding from an image. To support this task, MAIRA-2 introduces a custom dataset, explicitly designed to provide structured annotations aligning textual descriptions with spatial regions of interest in radiological images. This approach contrasts with conventional RRG models that generate unstructured free-text reports.

It is worth noting that the inference sets differ slightly across these models. Additionally, all these models leverage supplementary radiology information, such as lateral view scans, prior study reports, or both (as detailed in Appx. A.2), to enhance their performance in radiology-related tasks.

| Metric | RaDialog | MedVersa | MAIRA-2 | M4CXR | Libra (%) |
|---|---|---|---|---|---|
| **Lexical:** | | | | | |
| ROUGE-L | 31.6 | – | **38.4** | 28.5 | 36.7 (-4.4%) |
| BLEU-1 | 39.2 | – | 46.5 | 33.9 | **51.3** (10.3%) |
| BLEU-4 | 14.8 | 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%) |
| $RadCliQ_0(\downarrow)$ | – | 2.7 | **2.6** | – | 2.7 (-3.8%) |
| CheXbert vector | – | 46.4 | **50.6** | – | 46.9 (-7.3%) |
| *CheXpert-F1:* | | | | | |
| Micro-F1-14 | 39.2 | – | 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%) |

Table 7: Findings section 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. '–' indicates missing data. The percentage (%) indicates the improvement over the best existing model.

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.

### G.2 Discussion on Performance with non-LLM-based Models

To compare 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 and Lavie, 2005), and ROUGE-L (R-L) (Lin, 2004). For clinical metrics, we report CheXbert (Irvin et al., 2019), Precision (P), Recall (R), and $F_1$.

**Baseline** For performance evaluation, we compare our model with the following baselines: ST (Vinyals et al., 2015), ATT2IN (Rennie et al., 2017), ADAATT (Lu et al., 2017), TopDown (Anderson et al., 2018), R2Gen (Chen et al., 2020), R2GenCMN (Chen et al., 2021), $M^2$TR (Nooralahzadeh et al., 2021), CMCL (Liu et al., 2021a),PPKED (Liu et al., 2021b), AlignTransformer (You et al., 2021), CA (Liu et al., 2021c), LKBMA (Yang et al., 2022b), KnowMAT (Yang et al., 2022a), XPRONET (Wang et al., 2022), CMM-RL (Qin and Song, 2022), RAMT (Zhang et al., 2024b), CMCA (Song et al., 2022), KiUT (Huang et al., 2023), DCL (Li et al., 2023b), MMTN (Cao et al., 2023), METrans (Wang et al., 2023), ORGAN (Hou et al., 2023b), COMG (Gu et al., 2023), BioViL-T (Bannur et al., 2023), RGRG (Tanida et al., 2023), RECAP (Hou et al., 2023a), CvT2DistilGPT2 (Nicolson et al., 2023), VLCI (Chen et al., 2024a), TiBiX (Sanjeev et al., 2024), MedM2G (Zhan et al., 2024), MS-TF (Mei et al., 2024).

To ensure fairness, Libra also utilizes prior images, aligning with other models that leverage prior images or additional information. As demonstrated in Table 8, Libra, similar to other LLM-based models, consistently outperforms non-LLM-based models. This advantage is largely attributed to advancements in LLMs and visual instruction tuning (Liu et al., 2023), enabling multimodal large language models (MLLMs) to achieve superior performance in RRG tasks.

| Model | Lexical Metrics | | | | | | Clinical Metrics | | |
|---|---|---|---|---|---|---|---|---|---|
| | B-1 | B-2 | B-3 | B-4 | MTR | R-L | P | R | $F_1$ |
| ST‡ | 29.9 | 18.4 | 12.1 | 8.4 | 12.4 | 26.3 | 24.9 | 20.3 | 20.4 |
| ATT2IN‡ | 32.5 | 20.3 | 13.6 | 9.6 | 13.4 | 27.6 | 32.3 | 23.9 | 20.4 |
| ADAATT‡ | 29.9 | 18.5 | 12.4 | 8.8 | 11.8 | 26.6 | 26.8 | 18.6 | 18.1 |
| TopDown‡ | 31.7 | 19.5 | 13.0 | 9.2 | 12.8 | 26.7 | 32.0 | 23.1 | 23.8 |
| 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 |
| XPRONET | 34.4 | 21.5 | 14.6 | 10.5 | 13.8 | 27.9 | – | – | – |
| CMCL | 34.4 | 21.7 | 14.0 | 9.7 | 13.3 | 28.1 | – | – | – |
| PPKED | 36.0 | 22.4 | 14.9 | 10.6 | 14.9 | 28.4 | – | – | – |
| AlignTransformer | 37.8 | 23.5 | 15.6 | 11.2 | 15.8 | 28.3 | – | – | – |
| CA | 35.0 | 21.9 | 15.2 | 10.9 | 15.1 | 28.3 | 35.2 | 29.8 | 30.3 |
| LKBMA | 38.6 | 23.7 | 15.7 | 11.1 | – | 27.4 | 42.0 | 33.9 | 35.2 |
| $M^2$TR | 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 |
| RAMT | 36.2 | 22.9 | 15.7 | 11.3 | 15.3 | 28.4 | 38.0 | 34.2 | 33.5 |
| 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 |
| MMTN | 37.9 | 23.8 | 15.9 | 11.6 | 16.1 | 28.3 | – | – | – |
| 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 | – | – | – |
| CvT2DistilGPT2 | 39.2 | 24.5 | 16.9 | 12.4 | 15.3 | 28.5 | 35.9 | 41.2 | 38.4 |
| RGRG | 37.3 | 24.9 | 17.5 | 12.6 | 16.8 | 26.4 | 46.1 | 47.5 | 44.7 |
| BioViL-T | – | – | – | 9.2 | – | 29.6 | – | – | 17.5 |
| VLCI | 40.0 | 24.5 | 16.5 | 11.9 | 15.0 | 28.0 | 48.9 | 34.0 | 40.1 |
| 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 |
| MS-TF | 43.6 | 27.5 | 18.4 | 12.9 | 17.7 | 30.5 | – | – | 41.1 |
| Libra | **51.3** | **38.0** | **30.0** | **24.5** | **48.9** | **36.7** | **59.7** | **52.5** | **55.9** |

Table 8: Findings Generation Performance of Libra and non-LLM-based Models. The best performances are highlighted in **bold**, and the second-best scores are underlined. ‡ denotes results from Chen et al. (2021), and '–' indicates missing data. These results are taken from the best performances reported in their original papers.

# H Additional Ablation Studies

## H.1 Impact of Temporal Information on Libra in RRG

Temporal information is embedded in paired images and referenced in the corresponding radiology reports, capturing changes over time through references to prior symptoms and their progression, as discussed in Appx. B.1. As shown in Table 5, **86%** of the test data includes prior images, providing a solid foundation for evaluating the impact of temporal information.

During training, Libra integrates the ability to perceive and utilise temporal information into its architecture. To evaluate whether Libra effectively leverage temporal information during inference, we assess its performance using prior images when available as references to determine their impact on the overall capability.

In Table 9, the inclusion of prior images substantially enhances Libra's performance across all

| Metric | Libra | |
|---|---|---|
| | **w/o** prior | **w/** prior (%) |
| **Lexical:** | | |
| ROUGE-L | 36.17 | 36.66 (+1.35%) |
| BLEU-1 | 51.20 | 51.25 (+0.10%) |
| BLEU-4 | 24.33 | 24.54 (+0.86%) |
| METEOR | 48.69 | 48.90 (+0.43%) |
| BERTScore | 61.94 | 62.50 (+0.90%) |
| $F1_{temp}$ | 32.72 | 35.34 (+8.00%) |
| **Clinical:** | | |
| RadGraph-F1 | 32.42 | 32.87(+1.39%) |
| $RG_{ER}$ | 36.92 | 37.57(+1.76%) |
| $RadCliQ_0(\downarrow)$ | 2.76 | 2.72 (+1.45%) |
| CheXbert vector | 46.31 | 46.85 (+1.17%) |
| *CheXpert-F1:* | | |
| Micro-F1-14 | 55.25 | 55.87 (+1.12%) |
| Macro-F1-14 | 40.15 | 40.38(+0.57%) |
| Micro-F1-5 | 58.93 | 60.07(+1.93%) |
| Macro-F1-5 | 52.61 | 53.75(+2.17%) |

Table 9: Ablation results for Libra without **(w/o)** and with **(w/)** the prior image. Values in (%) indicate the percentage improvement.

| 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%) |

Table 10: 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*.

metrics. Notably, clinical scores exhibit greater improvements compared to lexical scores, underscoring the importance of temporal information in generating high-quality medical reports beyond merely improving linguistic fluency.

The $F1_{temp}$ score shows the most substantial improvement, with an increase of **8%**, highlighting Libra's capability to effectively leverage temporal changes provided by prior images. These results validate the role of temporal information in enhancing the quality of the generated *Findings* section and improving Libra's overall performance.

### H.2 Impact of the Temporal Alignment Connector under General-Domain Pre-trained Models

Domain-specific pre-trained models (i.e., RAD-DINO (Pérez-García et al., 2024) and Meditron (Chen et al., 2023c)) inherently incorporate domain-specific knowledge, such as phrasing conventions, pronoun usage, and even temporal information embedded in the training corpus. To isolate the structural impact of TAC, 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 replicated the first ablation setup from Sec. 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 10, we then conducted ablation studies by sequentially removing different components from the model, including the Temporal Fusion Module (TFM), Layerwise Feature Extractor (LFE), Prior Image Prefix Bias (PIPB), 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 feature with the LLM's hidden dimensions. Notably, without TFM, the model cannot process 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 using domain-specific models, as presented in Table 2. Removing any TAC submodule led to declines across all metrics. Specifically, removing TFM caused a notable drop in the $F1_{temp}$ score (↓>4%), emphasising its role in capturing temporal information. The absence of LFE significantly reduced RadGraph-related scores, demonstrating its importance for detailed image feature extraction. PIPB removal primarily impacted clinical metrics, while removing the entire TAC resulted in substantial declines across all metrics. These findings reaffirm the critical role of TAC in integrating image details and temporal information effectively.

### H.3 Impact of the Temporal Alignment Connector After the Second-Stage Fine-tuning

To further evaluate the impact of the Temporal Alignment Connector (TAC) on Libra's performance, we followed the setup of the first ablation study in Sec. 4. After the first stage of alignment,

| Metric | *Libra-2* | w/o TFM | w/o LFE | w/o PIPB | w/o TAC |
|---|---|---|---|---|---|
| **Lexical:** | | | | | |
| ROUGE-L | 35.31 | 35.16 (-0.42%) | 35.09 (-0.64%) | 35.23 (-0.23%) | 34.41 (-2.55%) |
| BLEU-1 | 49.92 | 49.44 (-0.97%) | 49.47 (-0.90%) | 49.75 (-0.34%) | 48.61 (-2.63%) |
| BLEU-4 | 23.05 | 22.67 (-1.66%) | 22.65 (-1.75%) | 22.97 (-0.35%) | 21.51 (-6.70%) |
| METEOR | 47.99 | 47.69 (-0.62%) | 47.62 (-0.77%) | 47.84 (-0.31%) | 46.95 (-2.16%) |
| BERTScore | 61.28 | 61.13 (-0.24%) | 61.07 (-0.34%) | 61.21 (-0.12%) | 60.60 (-1.12%) |
| $F1_{temp}$ | 33.52 | 33.10 (-1.27%) | 33.25 (-0.79%) | 33.49 (-0.09%) | 32.73 (-2.36%) |
| **Clinical:** | | | | | |
| RadGraph-F1 | 30.77 | 30.55 (-0.72%) | 30.43 (-1.10%) | 30.65 (-0.40%) | 30.07 (-2.27%) |
| $RG_{ER}$ | 35.44 | 35.16 (-0.79%) | 35.05 (-1.10%) | 35.29 (-0.42%) | 34.55 (-2.51%) |
| $RadCliQ_0$ (↓) | 2.83 | 2.84 (-0.35%) | 2.84 (-0.35%) | 2.85 (-0.71%) | 2.85 (-0.71%) |
| CheXbert vector | 45.32 | 45.08 (-0.53%) | 44.97 (-0.77%) | 45.27 (-0.11%) | 44.73 (-1.30%) |
| *CheXpert-F1:* | | | | | |
| Micro-F1-14 | 54.11 | 53.73 (-0.70%) | 53.70 (-0.76%) | 54.00 (-0.20%) | 53.41 (-1.30%) |
| Macro-F1-14 | 37.16 | 36.74 (-1.13%) | 36.78 (-1.02%) | 36.79 (-1.00%) | 36.64 (-1.40%) |
| Micro-F1-5 | 58.76 | 58.10 (-1.12%) | 58.32 (-0.75%) | 58.36 (-0.68%) | 57.65 (-1.89%) |
| Macro-F1-5 | 51.99 | 51.16 (-1.60%) | 51.19 (-1.54%) | 51.27 (-1.38%) | 50.87 (-2.15%) |

Table 11: Results of ablation experiments for the Temporal Alignment Connector after the second stage. '↓' indicates that lower is better. Values in (%) indicate the percentage decrease compared with the *Libra-2*.

the model underwent a second stage of fine-tuning. This stage was designed to optimise the model's performance on the *Findings* section generation task by leveraging the aligned visual and textual features learned during the initial stage.

In this phase, we applied Low-Rank Adaptation (LoRA) Hu et al. (2021) to fine-tune the pre-trained LLM (Meditron Chen et al. (2023c)), while keeping the visual encoder (RAD-DINO Pérez-García et al. (2024)) and TAC weights frozen. The baseline for this experiment is *Libra-2* (in Table 11), which is derived from *Libra-1* (in Table 2) after undergoing LoRA fine-tuning.

We conducted ablation studies by progressively removing different TAC components, including TFM, LFE, the Prior Image Prefix Bias (PIPB), and the entire TAC. Results consistently showed declines across all metrics compared to *Libra-2*, mirroring the trends observed in Sec. 4. This reinforces that the performance improvements brought by TAC are stable and unaffected by changes in training stages. It further confirms that TAC has embedded the capability to process temporal information within the model.

### H.4 Robustness Evaluation of the Temporal Alignment Connector

To evaluate the robustness of the Temporal Alignment Connector (TAC), we introduced an additional round of LoRA fine-tuning to induce overtraining. Following the setup in Appx. H.3, after integrating the first LoRA weights, a new set of LoRA adapters was reinitialised for the LLM and trained for one epoch under the same second-stage fine-tuning configuration. The baseline for this experiment is *Libra-3* (as shown in Table 12), which is derived from *Libra-2* (illustrated in Table 11) following this additional fine-tuning step.

The results reveal that, compared to *Libra-2*, *Libra-3* exhibits minimal changes in lexical scores, while clinical scores decline due to overfitting caused by the additional fine-tuning. Notably, the CheXpert (Smit et al., 2020a) (Macro-F1-[5/14]) scores exhibit the most influential reduction.

Despite this decline, ablation studies confirm that TAC's performance improvements remain robust, unaffected by variations in training strategies. This resilience stems from TAC's ability to capture and retain temporal image representations during the initial training phase, which are preserved through subsequent fine-tuning.

These findings underscore TAC's reliability as a critical component for temporal information processing in RRG tasks. It ensures stability even under diverse training conditions.

### H.5 Impact of Radiology-Specific Pre-trained Models on Libra

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 baseline setup of the previous ablation study in Sec. 4). Then we replaced the image encoder and

| Metric | *Libra-3* | w/o TFM | w/o LFE | w/o PIPB | w/o TAC |
|---|---|---|---|---|---|
| **Lexical:** | | | | | |
| ROUGE-L | 35.58 | 35.53 (-0.14%) | 35.51 (-0.21%) | 35.55 (-0.08%) | 35.28 (-0.86%) |
| BLEU-1 | 49.54 | 49.38 (-0.32%) | 49.39 (-0.30%) | 49.48 (-0.11%) | 49.10 (-0.88%) |
| BLEU-4 | 23.61 | 23.48 (-0.55%) | 23.47 (-0.58%) | 23.58 (-0.12%) | 23.07 (-2.28%) |
| METEOR | 47.61 | 47.51 (-0.21%) | 47.49 (-0.26%) | 47.56 (-0.10%) | 47.26 (-0.73%) |
| BERTScore | 61.54 | 61.49 (-0.08%) | 61.47 (-0.11%) | 61.52 (-0.04%) | 61.31 (-0.37%) |
| $F1_{temp}$ | 33.51 | 33.37 (-0.42%) | 33.42 (-0.27%) | 33.50 (-0.03%) | 33.24 (-0.79%) |
| **Clinical:** | | | | | |
| RadGraph-F1 | 29.82 | 29.75 (-0.24%) | 29.71 (-0.37%) | 29.78 (-0.13%) | 29.59 (-0.76%) |
| $RG_{ER}$ | 35.60 | 35.51 (-0.26%) | 35.47 (-0.37%) | 35.55 (-0.14%) | 35.30 (-0.84%) |
| $RadCliQ_0$ (↓) | 2.91 | 2.92 (-0.34%) | 2.92 (-0.34%) | 2.91 ( – ) | 2.93 (-0.68%) |
| CheXbert vector | 44.77 | 44.69 (-0.18%) | 44.65 (-0.26%) | 44.75 (-0.04%) | 44.57 (-0.45%) |
| *CheXpert-F1:* | | | | | |
| Micro-F1-14 | 52.45 | 52.33 (-0.23%) | 52.32 (-0.25%) | 52.41 (-0.08%) | 52.22 (-0.44%) |
| Macro-F1-14 | 30.77 | 30.65 (-0.38%) | 30.66 (-0.34%) | 30.67 (-0.33%) | 30.63 (-0.47%) |
| Micro-F1-5 | 54.42 | 54.22 (-0.38%) | 54.28 (-0.25%) | 54.30 (-0.23%) | 54.08 (-0.63%) |
| Macro-F1-5 | 44.58 | 44.34 (-0.54%) | 44.35 (-0.52%) | 44.37 (-0.46%) | 44.26 (-0.72%) |

Table 12: Results of ablation experiments for the Temporal Alignment Connector with additional LoRA fine-tuning after the second stage. '↓' indicates that lower is better. Values in (%) indicate the percentage decrease compared with the *Libra-3*.

| 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%) |

Table 13: 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*.

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* (in Table 10).

As shown in Table 13, 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.

### H.6 Incremental Component Analysis

We conducted an incremental study to evaluate the effectiveness of each component in Libra's architecture. Starting with a baseline model similar to LLaVA—comprising a pre-trained CLIP image encoder, a randomly initialised four-layer MLP adapter, and Vicuna-7B-v1.5 as the LLM—we trained the adapter on the *Findings* section generation task.

Improvements were introduced incrementally, as summarised in Table 14. First, we replaced the image encoder with DINOv2. Next, we incorporated the LFE (prefix module of TAC) and subsequently added the TFM (suffix module), completing the TAC connector. We then replaced the image encoder and LLM with RAD-DINO and Meditron,

| 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** |
| $RadCliQ_0(\downarrow)$ | 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** |

Table 14: 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 configuration. The best performances are highlighted in **bold**, and the second-best scores are underlined. '↓' indicates that lower is better.

respectively. The dataset for the first stage was expanded, and final fine-tuning was conducted for downstream tasks to produce Libra.

With each enhancement, the model's performance improved, demonstrating the critical role of each component. Notably, the addition of the TFM during the alignment stage provided the most significant improvement, showcasing its ability to capture temporal information, which is essential for the RRG task.

However, data expansion in the first stage led to improved lexical scores but a slight decline in clinical metrics, likely due to the VQA task's focus on fine-grained grounded information rather than holistic report generation, as mentioned in Sec. 4. This shift also affected the $F1_{temp}$ score, as temporal entities are often linked to specific symptoms. These declines were subsequently addressed through second-stage fine-tuning, resulting in overall improved performance.

**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 prior images, in Table 15.

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

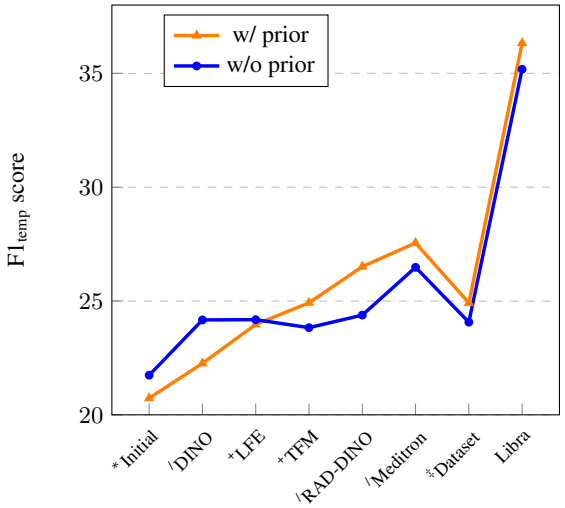


Table 15: Results of ablation experiments for Libra on the $F1_{temp}$ score. Among 2,461 test samples, 2,117 include a prior image, while 344 do not.

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 Sec. 5.

# I Heatmap Analysis and Temporal Feature Representation

The heatmap in Figure 4 corresponds to the example in ***(a)*** of Figure 3, where no prior image was used as a reference. It illustrates the clear differences in feature representations across layers of the RAD-DINO (Pérez-García et al., 2024) image

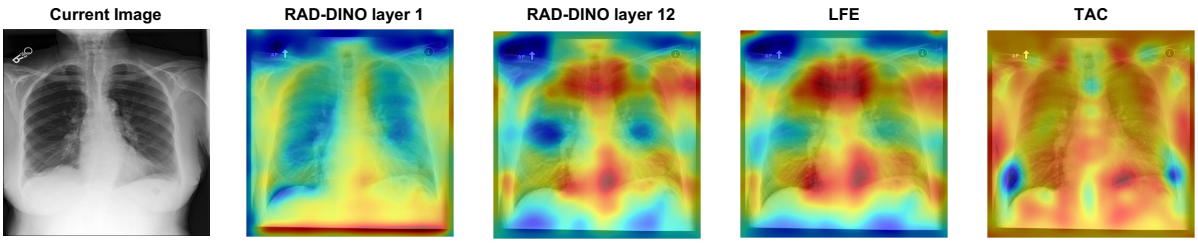


Figure 4: Heat map visualisation of image representations from different image encoder layers and the Temporal Alignment Connector (TAC), up-sampled using a Gaussian filter. Warm colours (red, yellow) indicate regions with higher weight allocations in the intermediate outputs of the "hidden-state" within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

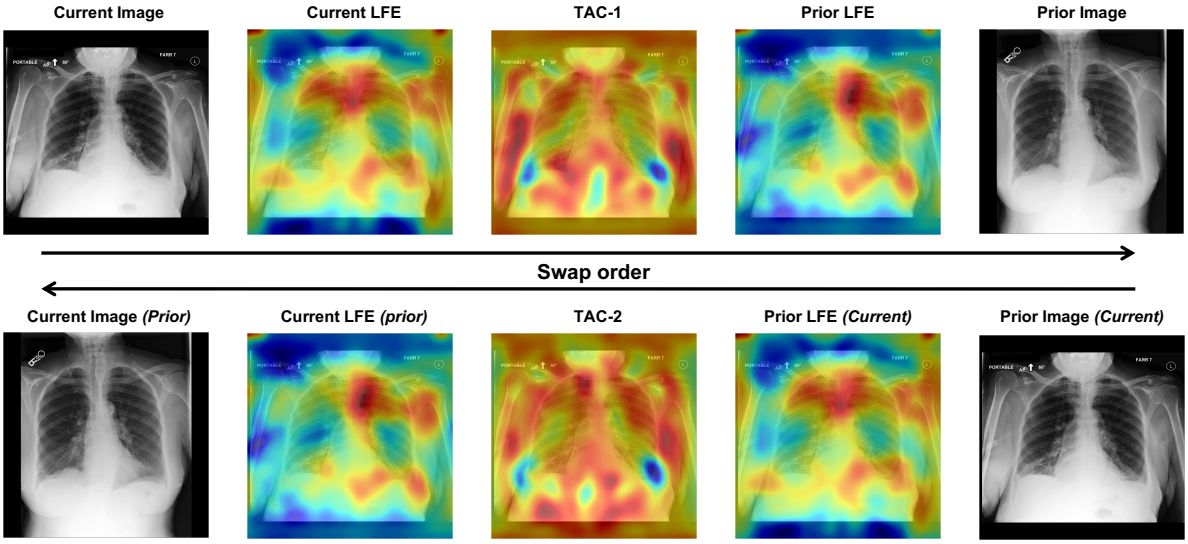


Figure 5: Heat map visualisation of image representations from the Temporal Alignment Connector (TAC), up-sampled using a Gaussian filter. The arrows ('→') represent the direction of temporal information, pointing from the prior image to the true current image. Warm colours (red, yellow) indicate regions 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.

encoder. The shallow layers primarily capture the overall lung structure, while the deeper layers focus on specific disease regions.

After passing through the Layerwise Feature Extractor (LFE), the image feature representations assign higher weights to larger symptom regions, achieving finer granularity. Following the Temporal Alignment Connector (TAC), the model integrates the weighted dummy prior image, producing a uniform feature distribution that reflects temporal information. This indicates no significant changes compared to the prior study and facilitates smoother image feature representations for downstream text generation by the LLM.

The heatmap in Figure 5 corresponds to the example in ***(b)*** of Figure 3, where a prior image is provided. After processing through the LFE, the model captures fine-grained feature representations in symptom areas. When processed by the TAC, these features are integrated with the differences between the two images, effectively reflecting temporal information, as demonstrated in TAC-1 (top) of Figure 5.

When the image order is swapped, treating the prior image as the current image, the LFE output remains unchanged. However, comparing TAC-2 (bottom of Figure 5) and TAC-1 outputs reveals significant differences in lung feature representations. This highlights the model's directional temporal perception and confirms that the TAC module effectively encodes temporal information from different time points, while the LFE focuses solely on image features without temporal encoding.

This behaviour aligns with the design of the TAC, where residual connections prioritise the current image as the main modality and the prior image as the auxiliary. Swapping the image order changes the main modality, altering the temporal state of symptoms in the generated report, such as reversing descriptions from "improving" to "worsening," as discussed in Sec. 5.

## J Extended Discussion on Limitations

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** Currently, our model captures temporal information through paired image comparisons and 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.

**Sparse Temporal Data Challenges** In cases where prior scans are unavailable or minimally informative (e.g., taken within a short interval), our "dummy prior image" provides a workaround. However, the model's ability to interpret and generate meaningful outputs under these constraints may still be limited. Future research could focus on synthesising or imputing temporal context to enhance performance under these constraints.

**Computational Complexity** The use of temporal alignment mechanisms and multi-layer feature integration increases computational demands, posing challenges for deployment in resource-constrained environments. Future optimisation efforts should focus on reducing computational overhead while maintaining performance.

**Generalisability Across Modalities and Datasets** Our study is limited to frontal-view chest X-rays and the MIMIC-CXR dataset (Johnson et al., 2019b). The applicability of our approach to other imaging modalities (e.g., CT, MRI) and datasets (e.g., CheXpert (Irvin et al., 2019), PadChest (Bustos et al., 2020)) remains unexplored. Future studies should assess the model's generalisability to a broader range of datasets and imaging contexts.

Based on the identified limitations, we outline the following directions:

- Develop frameworks for integrating multiple prior scans with dynamic temporal reasoning to better capture longitudinal changes.
- Expand the model to incorporate multi-modal imaging and textual data for more comprehensive diagnostic insights.
- Investigate the integration of diverse temporal data sources, such as electronic health records (EHRs), to enhance clinical applicability.
- Exploring lightweight model architectures for faster inference while maintaining high performance.

These advancements aim to address the current limitations while broadening the applicability of temporal-aware multimodal models in radiology and other clinical domains.