S2D-Align: Shallow-to-Deep Auxiliary Learning for Anatomically-Grounded Radiology Report Generation

Jiechao Gao[1,*], **Chang Liu**[2], **Yuangang Li**[3]
[1]Stanford University
[2]University of Science and Technology of China
[3]University of California, Irvine
jiechao@stanford.edu, christzhaung@gmail.com, yuanganl@uci.edu

Abstract

Radiology Report Generation (RRG) aims to automatically generate diagnostic reports from radiology images. To achieve this, existing methods have leveraged the powerful cross-modal generation capabilities of Multimodal Large Language Models (MLLMs), primarily focusing on optimizing cross-modal alignment between radiographs and reports through Supervised Fine-Tuning (SFT). However, by only performing instance-level alignment with the image-text pairs, the standard SFT paradigm fails to establish anatomically-grounded alignment, where the templated nature of reports often leads to sub-optimal generation quality. To address this, we propose S2D-ALIGN, a novel SFT paradigm that establishes anatomically-grounded alignment by leveraging auxiliary signals of varying granularities. S2D-ALIGN implements a shallow-to-deep strategy, progressively enriching the alignment process: it begins with the coarse radiograph-report pairing, then introduces reference reports for instance-level guidance, and ultimately utilizes key phrases to ground the generation in specific anatomical details. To bridge the different alignment stages, we introduce a memory-based adapter that empowers feature sharing, thereby integrating coarse and fine-grained guidance. For evaluation, we conduct experiments on the public MIMIC-CXR and IU X-RAY benchmarks, where S2D-ALIGN achieves state-of-the-art performance compared to existing methods. Ablation studies validate the effectiveness of our multi-stage, auxiliary-guided approach, highlighting a promising direction for enhancing grounding capabilities in complex, multi-modal generation tasks.

Introduction

Medical imaging, such as X-rays and Computed Tomography (CT), serves as an indispensable non-invasive tool in modern diagnostics, offering a crucial way to visualize the internal structures of human body conditions. Following the interpretation of these images, radiologists are required to record detailed diagnostic reports that translate complex visual findings into precise medical language, forming a critical basis for subsequent clinical decision-making. This manual process, however, is not only time-consuming but also susceptible to errors and omissions, particularly for less experienced radiologists, which can potentially degrade the quality of patient care. To mitigate these challenges, the task of Radiology Report Generation (RRG) has been motivated by recent studies (Jing, Xie, and Xing 2018a; Li et al. 2018; Chen et al. 2020b; Liu et al. 2021a; Chen et al. 2021a; Qin and Song 2022a), aiming to develop automatic solutions to alleviate the workload of radiologists, where this research direction has raised great attention from the communities of both artificial intelligence and clinical medicine.

Recent breakthroughs in Large Language Models (LLMs) (Touvron et al. 2023) have motivated Multimodal Large Language Models (MLLMs) (Zhu et al. 2023; Liu et al. 2023) as the cornerstone for RRG, effectively overcoming the alignment challenges inherent in earlier methods (Liu et al. 2023) trained from scratch on limited datasets. Adapting these general MLLMs for the medical domain primarily involves two competing strategies, i.e., In-Context Learning (ICL) and Supervised Fine-Tuning (SFT). ICL methods (Yan et al. 2023), which keep the LLM parameters frozen, typically rely on external annotators like RadGraph (Jain et al. 2021) to convert visual information into structured text (e.g., entities and relations), upon which few-shot demonstrations guide the generation. However, their performance is highly sensitive to the quality of these text-based representations and the choice of demonstration examples, limiting their robustness in complex clinical scenarios. Consequently, SFT has emerged as the dominant paradigm, establishing end-to-end alignment by directly fine-tuning the MLLM on radiograph-report pairs (Liu et al. 2024; Wang et al. 2025; Hyland et al. 2023; Tu et al. 2023). Despite its prevalence, the standard SFT framework faces a critical bottleneck, where it performs alignment only at a coarse granularity between the entire image and its corresponding report. This coarse-grained approach, confounded by the templated and often redundant nature of radiology reports, fails to establish precise correspondence between specific pathological findings and their anatomical locations. This deficiency in alignment granularity directly undermines the factual correctness and clinical reliability of the generated reports. Architecturally, this limitation is often exacerbated by the use of simple projection layers that bridge the visual encoder and the LLM, which are insufficient for learning fine-grained, region-to-text mappings for RRG. Therefore, developing an effective method for *anatomically-grounded alignment* has become a critical challenge for trustworthy RRG, where this

*Corresponding author

pivotal problem motivates our work in this paper.

To address this critical challenge, we introduce S2D-ALIGN, a novel fine-tuning paradigm designed to explicitly establish anatomically-grounded alignment. At its core, we propose **Progressive Anatomical Grounding (PAG)**, a shallow-to-deep SFT strategy that systematically enriches the alignment process by leveraging auxiliary signals of varying granularities. This multi-stage process begins with coarse-grained radiograph-report alignment, then incorporates reference reports for enhanced contextual understanding, and culminates in fine-grained grounding using clinically-relevant key phrases to connect text to specific anatomical regions. To unify the learning signals across the diverse stages of PAG, we introduce the **Shallow-to-Deep Memory Adapter (SMA)**, a lightweight yet effective memory-based adapter that facilitates feature sharing and integrates coarse- and fine-grained guidance into a cohesive representation. Our extensive experiments on the IU X-RAY and MIMIC-CXR benchmarks demonstrate that S2D-ALIGN achieves new state-of-the-art performance compared to prevailing methods. Generally speaking, the contributions of S2D-ALIGN are threefold:

- We propose **Progressive Anatomical Grounding (PAG)**, an innovative multi-stage SFT framework that explicitly targets anatomically-grounded RRG;
- We design the **Shallow-to-Deep Memory Adapter (SMA)**, an effective module that enable multi-grained feature sharing during the fine-tuning of MLLMs;
- We conduct comprehensive experiments that not only validate the superiority of S2D-ALIGN, but also highlight a promising direction for building more factually reliable and clinically trustworthy generative models.

Related Work

The advent of deep learning has catalyzed a significant paradigm shift in RRG over the last decade. Foundational approaches (Jing, Xie, and Xing 2018a; Li et al. 2018; Liu et al. 2021a,c; Nicolson, Dowling, and Koopman 2022; Huang, Zhang, and Zhang 2023; Tanida et al. 2023; Liu, Tian, and Song 2024; Jin et al. 2024) established the encoder-decoder framework, typically by training task-specific neural networks on benchmark datasets (Demner-Fushman et al. 2016; Johnson et al. 2019). These models primarily focused on enhancing cross-modal alignment to improve report quality, employing techniques such as memory networks (Chen et al. 2020b, 2021a), attention mechanisms (Liu et al. 2021a), reinforcement learning (Qin and Song 2022a), etc. However, these methods, trained from scratch on limited-scale medical datasets, were fundamentally constrained in model capacity, limiting their applicability to complex, real-world clinical scenarios. The emergence of Large Language Models (LLMs), pre-trained on massive text corpora, has introduced powerful text generation capabilities, motivating a new research direction to overcome the limitations of earlier methods. Consequently, the dominant paradigm has shifted towards adapting these models for RRG by aligning a visual encoder with an LLM and performing Supervised Fine-Tuning (SFT) on radiograph-report pairs (Hyland et al. 2023; Tu et al. 2023; Liu et al. 2024; Wang et al. 2025). Nevertheless, this standard SFT paradigm conducts at an instance-level of alignment, failing to establish the fine-grained mappings between specific visual findings and their textual descriptions necessary for anatomical grounding. Among the most relevant works, LLM-RG4 (Wang et al. 2025) attempts to address this by introducing an adaptive token fusion module and a token-level loss weighting strategy to prioritize descriptions of local regions. Yet, its learning process is still fundamentally constrained by instance-level data pairs, lacking explicit anatomical guidance. In contrast, our proposed S2D-ALIGN directly tackles this challenge by injecting explicit, multi-grained anatomical signals—such as key phrases and their corresponding visual regions—into the SFT process to progressively achieve anatomically-grounded alignment.

Methodology

In this section, we detail the architecture and training methodology of S2D-ALIGN. As illustrated in Figure 1, our framework is built upon three core modules, i.e., a frozen medical visual encoder ($\mathcal{E}_v$), our proposed **Shallow-to-Deep Memory Adapter (SMA)**, and a Large Language Model (LLM) decoder ($\mathcal{G}_{LLM}$). The fine-tuning of these modules is organized by our central contribution, the **Progressive Anatomical Grounding (PAG)** strategy, which leverages auxiliary signals to guide the model towards anatomically-grounded alignment.

The PAG strategy formalizes the fine-tuning as a three-stage curriculum. At each stage i, the model is conditioned on a progressively enriched multi-modal context $C^{(i)}$. Let I be the input radiograph, R_{ref} be the reference report, and K be the set of key phrases. We define the contexts as follows:

- **Stage 1 (Coarse Alignment):** The context contains only the visual information distilled by the SMA.

$$C^{(1)} \triangleq \text{SMA}_v(\mathcal{E}_v(I)) \tag{1}$$

- **Stage 2 (Contextual Enhancement):** The context is augmented with features from the reference report.

$$C^{(2)} \triangleq \texttt{concat}\left(C^{(1)}, \text{SMA}_\text{t}(\mathcal{E}_\text{text}(R_{ref}))\right) \tag{2}$$

- **Stage 3 (Fine-grained Grounding):** The context is further enriched with key phrase features.

$$C^{(3)} \triangleq \texttt{concat}\left(C^{(2)}, \text{SMA}_\text{p}(\mathcal{E}_\text{text}(K))\right) \tag{3}$$

where SMA_v, SMA_t, and SMA_p are the memory-based adapter modules for vision, reference reports, and key phrases, respectively, and $\texttt{concat}(\cdot)$ denotes the operation of concatenation along the channel dimension. Given this formulation, the training objective for the i-th stage of PAG is to minimize the auto-regressive cross-entropy loss over the ground-truth report R_{gt}:

$$\mathcal{L}^i_{\text{PAG}} = -\sum_{t=1}^{|R_{gt}|} \log p_{\Theta_i}\left(w_t | w_{<t}, C^{(i)}\right) \tag{4}$$

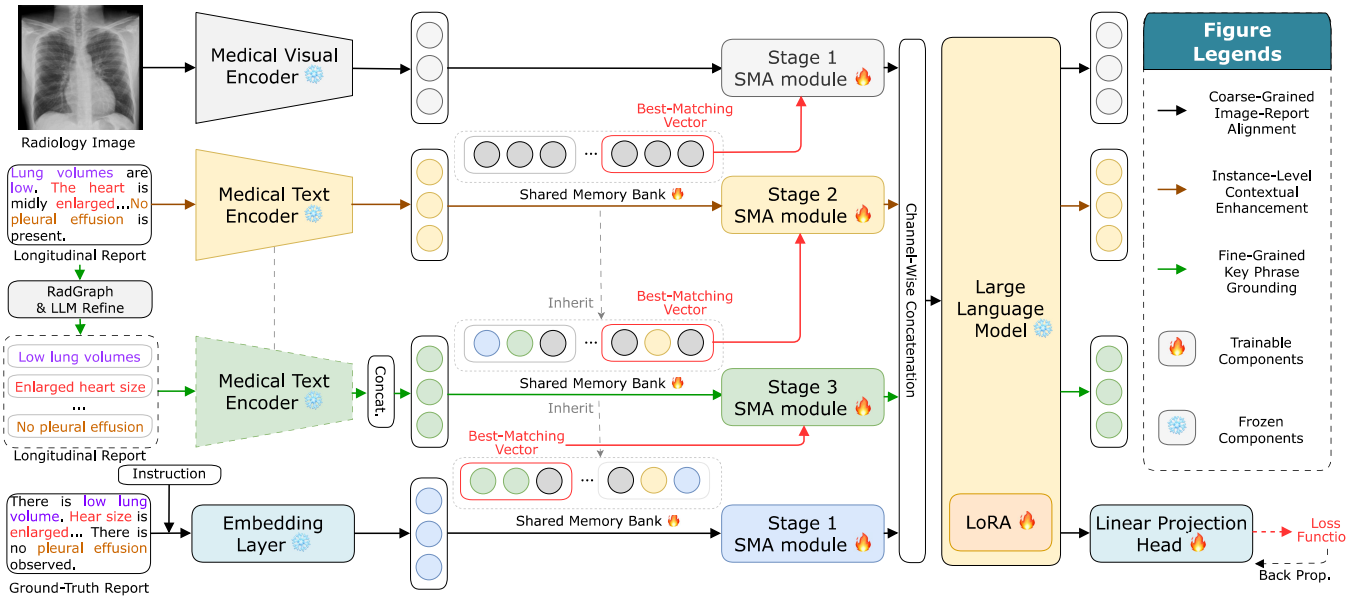

Figure 1: Overview of S2D-ALIGN, with the Progressive Anatomical Grounding (PAG) and Shallow-to-Deep Memory Adapter (SMA) modules as its core components. Herein, we use the same medical text encoders to convert reference reports or key phrases into embeddings, and adopt a shared memory bank inherited from earlier stages to later ones throughout PAG.

where w_t is the t-th token of R_{gt}, and p_{Θ_i} is the probability predicted by the model with parameters Θ_i trainable at stage i. During inference, we discard the auxiliary signals ($C^{(i)}, i \geq 2$) and generate the report R conditioned solely on the visual context $C^{(1)}$, which is formally expressed as:

$$R = \mathcal{G}_{\text{LLM}}\left(\cdot | C^{(1)}\right) \tag{5}$$

This training-inference asymmetry is a key design principle in this work, enabling the model to learn from a multi-modal context while maintaining the efficiency of a standard RRG pipeline. In the subsequent sections, we first introduce the visual feature extraction process, then illustrate the architecture of the SMA, and finally detail the PAG strategy.

Visual Encoder

The visual backbone of our framework is a pre-trained medical visual encoder, $\mathcal{E}_v$, which aims to encode an input radiology image $I \in \mathbb{R}^{H\times W\times C}$ into a sequence of feature embeddings. We adopt the Vision Transformer (ViT) architecture (Dosovitskiy et al. 2021), which processes the image by partitioning it into a sequence of non-overlapping patches. Latter, these patches are then embedded, incorporating positional information, to produce a sequence of patch-level feature vectors $V = \{v_1, v_2, \ldots, v_N\}$, which are then adopted different stages throughout the entire process of PAG.

Shallow-to-Deep Memory Adapter (SMA)

As is noted above, a pivotal component in standard MLLMs is the connector module that bridges the visual encoder and the LLM. Existing MLLMs typically employ simple connector designs like an MLP or with a further integration of Q-Former (Li et al. 2023), to project visual features into the embedding space of the LLM. However, these approaches are insufficient for establishing the anatomically-grounded alignment in the context of RRG, due to two main reasons, where the insufficient capabilities of MLPs and failure of feature sharing across modalities to capture complex relationships and complementary information. To address these shortcomings, we introduce the **Shallow-to-Deep Memory Adapter (SMA)**, a novel and efficient module designed to foster deep and interactive cross-modal alignment. Unlike conventional connectors, the SMA operates based on a multi-head cross-attention mechanism, along with a *memory bank* as a collection of N_{mem} learnable query vectors, denoted as $Q_{\text{mem}} \in \mathbb{R}^{N_{\text{mem}}\times D_v}$. During training, these memory queries dynamically interact with V from the encoder, adaptively attending to and distilling the most salient visual information into a compact representation. More importantly, the same memory bank is shared across all distinct alignment stages of PAG, enabling implicit feature sharing and compelling the adapter to learn a holistic and highly informative representation. Given the feature F_{aux} of the auxiliary signals (such as features of radiograph, reference report, and key phrases), this process is formally expressed as:

$$F_{\text{mem}} = \texttt{CrossAttn}(Q_{\text{mem}}, F_{aux}, F_{aux}; \Theta_{\text{SMA}}) \tag{6}$$

where $F_{\text{mem}} \in \mathbb{R}^{N_{mem}\times D_{aux}}$ is the resulting memory-enhanced feature that is fed to the LLM, and D_{aux} varies according to different modalities.

Progressive Anatomical Grounding (PAG)

The core contribution of our approach is the Progressive Anatomical Grounding (PAG) strategy, which addresses the limitations of standard instance-level SFT. PAG is conceptually motivated by Curriculum Learning (CL) (Bengio et al.

2009), a training paradigm that advocates for presenting easier examples to a model before progressively introducing more complex ones, thereby improving convergence and generalization. However, a direct application of CL to RRG is challenging, as defining a meaningful difficulty metric for radiograph-report pairs is non-trivial, yet simple heuristics, such as report length or the number of findings, often fail to capture trustworthy clinical complexity and the required level of grounding. To circumvent this challenge, PAG redefines the notion of difficulty not at the level of individual data samples, but the alignment task itself, which naturally fits the SFT nature of existing MLLMs. In doing so, it performs a multi-stage curriculum that progressively increases the required alignment granularity, guiding the model from learning instance-level semantics to anatomically-grounded descriptions. This is achieved by gradually introducing auxiliary textual signals of varying granularities across three individual stages, as detailed subsequently.

Stage 1: Coarse-Grained Image-Report Alignment. This initial stage of PAG establishes the basic alignment between the visual and textual modalities. Given an input radiograph $I \in \mathbb{R}^{H\times W\times C}$ and its ground-truth report $R_{gt} = \{w_1, w_2, \ldots, w_{|R_{gt}|}\}$, the forward pass proceeds as follows. First, the frozen medical visual encoder $\mathcal{E}_v$ processes the image to extract a sequence of patch-level feature embeddings $V \in \mathbb{R}^{N\times D_v}$. Next, these visual features V are fed into our lightweight SMA, whose parameters Θ_{SMA} are the sole target to update in this stage. The SMA comprises a series of learnable memory vectors, denoted as $Q_{\text{mem}} \in \mathbb{R}^{N_{\text{mem}}\times D_v}$, and levearges Q_{mem} to adaptively distill the visual representation V into a memory-enhanced one $V_{\text{mem}} \in \mathbb{R}^{N_{\text{mem}}\times D_v}$ via a cross-attention mechanism, written as:

$$V_{\text{mem}} = \text{SMA}_v(Q_{\text{mem}}, V, V; \Theta_{\text{SMA}_v}) \tag{7}$$

V_{mem} is then concatenated with the token embedding $E_{<t} = \text{Embed}(w_{<t})$ of R_{gt} along the sequence dimension, with the resulting representation serving as the visual context for the LLM $\mathcal{G}_{\text{LLM}}$, which eventually predicts the probability distribution over the vocabulary for the next token w_t, conditioned on V_{mem} and $w_{<t}$, written as:

$$p(w_t|w_{<t}, V; \Theta_{\text{SMA}_v}) = \mathcal{G}_{\text{LLM}}\left(w_{<t}, V_{\text{mem}}\right) \tag{8}$$

The training objective is to minimize the standard autoregressive Cross-Entropy loss (CE), which maximizes the likelihood of the ground-truth report, written as:

$$\mathcal{L}^1_{\text{PAG}} = -\sum_{t=1}^{|R_{gt}|} \log p(w_t|w_{<t}, V; \Theta_{\text{SMA}_v}) \tag{9}$$

By optimizing this objective, we exclusively update the parameters Θ_{SMA_v}, effectively aligning the feature space of the visual encoder with that of the LLM at an instance-level.

Stage 2: Instance-level Contextual Enhancement. To mitigate the ambiguity caused by the templated nature of radiology reports, this stage further introduces a reference report to enhance the instance-level context. To implement this, we leverage the inherent longitudinal nature of the MIMIC-CXR dataset. For a given radiograph-report pair (I, R_{gt}) from a specific patient study, we select the reference report R_{ref} from a different study of the same patient. This strategy is clinically motivated, where serial radiographs of the same individual normally share a high degree of anatomical correspondence, yet their reports often differ based on subtle but diagnostically critical interval changes. Therefore in this stage, the overall pipeline is then tasked to generate the correct report R_{gt} conditioned on both I and R_{ref}, forcing it to discover case-specific visual cues. To bridge R_{ref} with the input of I, we adopts an BERT-based text encoder model to convert R_{ref} into the corresponding representation $E_{\text{ref}} \in \mathbb{R}^{|R_{\text{ref}}|\times D_t}$, and utilizes a lightweight text adapter, architecturally identical to our SMA but with separate parameters (Θ_{SMA_t}), to project the text embedding E_{ref} into the shared feature space. Crucially, while the parameters of the text adapter are distinct, the concatenated features are eventually processed in the context of the same memory queries Q_{mem}, ensuring consistent feature integration. Then, the resulting representation is concatenated with V_{mem} to form the input of the LLM $\mathcal{G}_{\text{LLM}}$, which eventually predicts the probability distribution similar to that of Eq. 9, with the training objective $\mathcal{L}^2_{\text{PAG}}$ formulated by:

$$\mathcal{L}^2_{\text{PAG}} = -\sum_{t=1}^{|R_{gt}|} \log p(w_t|w_{<t}, V, R_{\text{ref}}; \Theta_{\text{SMA}_v}, \Theta_{\text{SMA}_t}) \tag{10}$$

Note that both SMA modules share the same memory vectors Q_{mem} in this stage, where the additional reference report guides the model to develop a more robust instance-level understanding by contrasting against similar cases.

Stage 3: Fine-grained Key Phrase Grounding. With solid instance-level visual understanding, this final stage aims to explicitly steer the model towards anatomically-grounded alignment. We first extract a set of clinically-relevant key phrases $K = \{k_1, k_2, \ldots, k_m\}$ from the ground-truth report R_{gt} using an entity extraction tool RadGraph (Jain et al. 2021). To ensure the grammatical coherence of the extracted entities, we first compose them into a short description if its corresponding relation is positive, then adopt an LLM to refine it into a more natural and clinically-relevant phrase.

With all training samples annotated, we randomly sample l key phrases from K and feed them to the LLM $\mathcal{G}_{\text{LLM}}$ for anotminal grounding. Particularly, we use the same medical text encoder as that in last stage to convert the key phrases into the corresponding representation $E_{\text{key}} \in \mathbb{R}^{l\times D_t}$ and use another SMA (SMA_p with parameters Θ_{SMA_p}) to map E_{key} into the same feature space as V and E_{ref}. Finally, we send the concatenation of V_{mem}, E_{ref}, and E_{key} to the LLM $\mathcal{G}_{\text{LLM}}$, with the training objective $\mathcal{L}^3_{\text{PAG}}$ formulated by:

$$\begin{aligned}\mathcal{L}^3_{\text{PAG}} = -\sum_{t=1}^{|R_{gt}|} \log p(w_t|w_{<t}, V_{\text{mem}}, E_{\text{ref}}, E_{\text{key}}; \\ \Theta_{\text{SMA}_v}, \Theta_{\text{SMA}_t}, \Theta_{\text{SMA}_p})\end{aligned} \tag{11}$$

By conducting this stage, the model is enforced to establish a more precise correspondence between visual regions

Model	NLG Metrics					CE Metrics		
	B@1	B@2	B@3	B@4	R-L	Precision	Recall	F1
Early Image Captioning Methods								
ST (Vinyals et al. 2015)	0.299	0.184	0.121	0.084	0.263	0.249	0.203	0.204
Att2In (Rennie et al. 2017)	0.325	0.203	0.136	0.096	0.276	0.322	0.239	0.249
AdaAtt (Lu et al. 2017)	0.299	0.185	0.124	0.088	0.266	0.268	0.186	0.181
TopDown (Anderson et al. 2018)	0.317	0.195	0.130	0.092	0.267	0.320	0.231	0.238
From-Scratch RRG Methods								
R2Gen (Chen et al. 2020b)	0.353	0.218	0.145	0.103	0.277	0.333	0.273	0.276
CA (Liu et al. 2021a)	0.350	0.219	0.152	0.109	0.283	-	-	-
CMCL (Liu et al. 2021c)	0.344	0.217	0.140	0.097	0.281	-	-	-
PPKED (Liu et al. 2021b)	0.360	0.224	0.149	0.106	0.284	-	-	-
R2GenCMN (Chen et al. 2021a)	0.353	0.218	0.148	0.106	0.278	0.334	0.275	0.278
R2GenRL (Qin and Song 2022a)	0.381	0.232	0.155	0.109	0.287	0.342	0.294	0.292
ITA (Wang et al. 2022)	0.395	0.253	0.170	0.121	0.284	-	-	-
WarmStart (Nicolson, Dowling, and Koopman 2022)	0.392	0.245	0.169	0.124	0.285	0.359	0.412	0.384
KiUT (Huang, Zhang, and Zhang 2023)	0.393	0.243	0.159	0.113	0.285	0.371	0.318	0.321
PromptMRG (Jin et al. 2024)	0.398	-	-	0.112	0.258	0.501	0.509	0.476
RGRG (Tanida et al. 2023)	0.373	0.249	<u>0.175</u>	0.126	0.264	0.461	0.475	0.447
Large Language Model-based RRG Methods								
XrayGPT (Thawkar et al. 2023)	0.128	0.045	0.014	0.004	0.111	-	-	-
Med-PaLM (Tu et al. 2023)	0.317	-	-	0.115	0.275	-	-	0.378
R2GenGPT (Wang et al. 2023)	0.396	-	-	0.113	0.273	0.506	0.414	0.456
EKAGen (Bu et al. 2024)	0.419	<u>0.258</u>	0.170	0.119	0.287	0.517	0.483	0.499
CheXAgent (Chen et al. 2024)	0.189	-	-	0.040	0.208	0.506	0.306	0.381
MAIRA-1 (Hyland et al. 2023)	0.392	-	-	<u>0.142</u>	0.289	-	-	0.553
R2-LLM (Liu et al. 2024)	0.402	-	-	0.128	0.291	0.465	0.482	0.473
InVERGe (Deria et al. 2024)	**0.425**	0.240	0.132	0.100	0.309	-	-	-
LLM-RG4 (Wang et al. 2025)	0.377	-	-	0.144	<u>0.318</u>	<u>0.583</u>	<u>0.593</u>	<u>0.588</u>
S2D-ALIGN (Ours)	<u>0.422</u>	**0.263**	**0.183**	**0.149**	**0.332**	**0.613**	**0.606**	**0.608**

Table 1: Comparison with state-of-the-art methods on the MIMIC-CXR benchmark (Johnson et al. 2019) with respect to standard NLG and CE metrics. The best and second best results are highlighted in bold and underlined, respectively.

and their textual descriptions, which is the cornerstone of trustworthy RRG. In inference, we discard all aforementioned auxiliary signals and perform RRG similar to standard MLLMs, resulting in the generated report R, where our experiments below demonstrate such paradigm effectively assists single radiograph-to-report generation without annotational guidance.

Experimental Setup

Datasets

We conduct our primary experiments on two benchmark datasets, i.e., IU X-RAY and MIMIC-CXR. IU X-RAY (Demner-Fushman et al. 2016) is collected by the Indiana University, where it serves one of the most widely adopted benchmarks for RRG, containing $7,470$ chest X-ray images and $3,955$ corresponding radiology reports. MIMIC-CXR (Johnson et al. 2019) is the largest publicly available dataset of chest X-ray radiographs and their corresponding free-text radiology reports, collected from the Beth Israel Deaconess Medical Center (BIDMC), where it contains $377,110$ image-report pairs from $65,379$ patients. Following the pre-processing pipeline of conventional studies (Chen et al. 2020b, 2021a; Liu et al. 2024; Wang et al. 2025), we extract the "Findings" section from each report for our analysis. We adhere to the official data split, consisting of $270,790$ training, $2,130$ validation, and $3,858$ testing samples. For data samples on MIMIC-CXR, we exclude samples lacking RadGraph annotations (Jain et al. 2021) to ensure that all key phrases are properly annotated.

Evaluation Metrics

Following standard practice (Li et al. 2018; Chen et al. 2020a, 2021b; Qin and Song 2022b), we first evaluate the generated reports using metrics like Natural Language Generation (NLG) and Clinical Efficacy (CE). Specifically, we employ BLEU-n (Papineni et al. 2002) (B@n, $n \in \{1, 4\}$) and ROUGE-L (Lin 2004) (R-L) for NLG assessment, and report precision, recall, and F1 scores for CE evaluation.

Implementation Details

For our implementation of the visual encoder $\mathcal{E}_v$, we leverage the pre-trained Rad-DINO (Pérez-García et al. 2024), and keep its parameters frozen throughout the fine-tuning process, in order to preserve its domain-specific visual representations learned from large-scale medical data. For the LLM decoder, we initialize the first stage of PAG with `Vicuna-7B-v1.5`, whose parameters are fixed during the first PAG stage, and fine-tuned via Low-Rank Adaptation (LoRA) (Hu et al. 2021). Herein, we apply LoRA to all linear layers of the transformer blocks with a rank of 16, a scaling factor of 16, and a dropout

Radiology Image

Ground-Truth Report: There is a new moderate left and small right **pleural effusion**. Right lower lobe **atelectasis** has slightly worsened. There is an indistinct haziness over the right lower **lung** field which may represent layering effusion. There is stable bilateral apical **pneumothoraces**. The **cardiomediastinal contour** is stable and demonstrates a mildly enlarged **heart**.

R2GenCMN (From-Scratch RRG Methods)

The **lung** volumes are low. The **heart** size is normal. There is a tiny right **pleural effusion**. Prosthetic valve is again noted with a right to moderate hiatal hernia. No signs of pneumonia edema. Bony structures are intact.

XrayGPT (SFT-based Methods, only Linear)

The radiology image is a frontal chest radiograph of an adult male. The image is oblique, with no abnormalities visible. The **lungs** are well aerated and the **heart** is in its normal position. There is no evidence of any infiltrates or consolidation.

LLM-RG4 (SFT-based Methods, Linear & LLM LoRA)

Frontal chest radiograph is presented. The lower lobe **atelectasis** is worsened. The chest radiograph shows moderate opacities in the right **lung**. The heart is mildly **enlarged**. **Cardiomediastinal contours** are clear and there is no **pleural effusion**.

Ours (PAG, Linear & LLM LoRA)

There is increased opacity in the right lower lobe, consistent with worsening **atelectasis**. Left and small right **pleural effusion** are seen. **Cardiomediastinal contour** are clear and stable. **Heart** size is mildly enlarged. Opacity is observed at the right **lung** base.

Figure 2: A case study selected from MIMIC-CXR, with medical concepts shared by the ground-truth and generated outputs highlighted in the same color. The categories and optimized parameters for the LLM-based methods are detailed in parentheses.

Methods	B@1	B@4	R-L
Early Image Captioning Methods			
ST (Vinyals et al. 2015)	0.216	0.066	0.306
Att2In (Rennie et al. 2017)	0.224	0.068	0.308
ADAATT (Lu et al. 2017)	0.220	0.068	0.308
CoATT (Jing, Xie, and Xing 2018b)	0.455	0.154	0.369
HRGR (Li et al. 2018)	0.438	0.151	0.322
CMAS-RL (Jing, Wang, and Xing 2019)	0.464	0.154	0.362
From-Scratch RRG Methods			
R2Gen (Chen et al. 2020b)	0.470	0.165	0.371
CA (Liu et al. 2021a)	0.492	0.169	0.381
CMCL (Liu et al. 2021c)	0.473	0.162	0.378
PPKED (Liu et al. 2021b)	0.483	0.168	0.376
R2GenCMN (Chen et al. 2021a)	0.475	0.170	0.375
R2GenRL (Qin and Song 2022a)	0.494	0.181	0.384
Large Language Model-based RRG Methods			
XrayGPT (7B) (Thawkar et al. 2023)	0.177	0.007	0.203
R2-LLM (14.2B)† (Liu et al. 2024)	0.499	0.184	0.390
Ours (7B)	**0.512**	**0.195**	**0.407**

Table 2: Comparison with state-of-the-art methods on IU X-RAY (Demner-Fushman et al. 2016) w.r.t. NLG metrics.

rate of 0.1, where the original LLM parameters are kept frozen, with only the LoRA adapter weights being updated during training. For the text encoder $\mathcal{E}_{\text{text}}$, we use `BiomedVLP-CXR-BERT-specialized` to encode the reference report and key phrases. For the SMA in different stages of PAG, it consists of a 8-head cross-attention layer followed by a 3-layer MLP and Layer Normalization (LN). The hyperparameters for S2D-ALIGN were tuned on the validation set, where we performed a grid search over key parameters to identify the optimal configuration

Results and Analysis

Comparison with State-of-the-Art Methods. As presented in the quantitative comparison of Table 1 and 2, we conduct a comprehensive comparison of our proposed S2D-ALIGN with a wide range of state-of-the-art methods on both the IU X-RAY (Demner-Fushman et al. 2016) and MIMIC-CXR (Johnson et al. 2019). Our proposed S2D-ALIGN establishes a new state-of-the-art on both benchmarks, achieving a F1 score of 0.608 on the CE metric. This superiority stems directly from our PAG strategy to progressively guide the model towards fine-grained radiograph-report alignment, which shows significantly lower performance on factual correctness (i.e., CE and RadGraph-based metrics) and reveals their failure to precisely map visual findings to precise textual descriptions. In contrast, the leading performance demonstrated by S2D-ALIGN provide direct evidence that it establishes *anatomically-grounded alignment* crucial for clinical reliability. Furthermore, our end-to-end paradigm, unified by the SMA design, overcomes the limitations of ICL (e.g., Yan et al. (Yan et al. 2023)) by avoiding reliance on potentially noisy intermediate textual representations, meanwhile enabling the models to discover more potential complementary information through feature sharing across modalities. These results validate that S2D-ALIGN is a more robust and factually reliable direction.

Case Study. We present a qualitative case study to intuitively compare the capabilities of representative RRG methods in Figure 2, with models from three categories, i.e., a from-scratch RRG method (R2GenCMN), an SFT-based MLLM with only the projector trained (XrayGPT), and an SFT-based MLLM that updates parameters of the LLM (LLM-RG4). As presented, R2GenCMN fails to capture key abnormalities like "atelectasis" and hallucinates

Method	B@1	B@4	R-L	F1
Evaluating the progressive nature of PAG				
Single-Stage (S1 only)	0.369	0.092	0.271	0.449
Joint Training (S1+S2+S3)	0.371	0.102	0.295	0.470
Reversed Order (S1→S3→S2)	0.355	0.077	0.301	0.527
Evaluating the contribution of each PAG stage				
w/o Fine-grained Grounding (S1→S2)	0.386	0.112	0.303	0.513
w/o Contextual Enhancement (S1→S3)	0.415	0.145	0.318	0.565
S2D-Align (Full, S1→S2→S3)	**0.422**	**0.149**	**0.332**	**0.608**

Table 3: Evaluation scores for the ablation studies of PAG on MIMIC-CXR, where "Si" denotes Stage *i*, "→" indicates progressive training, and "+" indicates joint training.

Connector Module	B@1	B@4	R-L	F1
MLP	0.387	0.104	0.283	0.473
MLP + Q-Former	0.368	0.098	0.264	0.394
SMA (MLP + MSA)	0.407	0.119	0.311	0.523
SMA (w/o Shared Memory)	0.401	0.136	0.320	0.559
SMA (Ours)	**0.422**	**0.149**	**0.332**	**0.608**

Table 4: Evaluation scores for the ablation studies of SMA on MIMIC-CXR, where "MLP" and "MLP + Q-Former" denote the linear projection and its combination with Q-Former (Li et al. 2023). "SMA (MLP + MSA)" and "SMA (w/o Shared Memory)" indicate the architecture without the memory bank and the sharing mechanism, respectively.

non-existent findings such as a "prosthetic valve", showcasing the limitations of training without leveraging pre-trained knowledge. XrayGPT produces a coarse-grained report with facutally incorrect diagnosis, i.e., "no abnormalities visible", and misses most pathological findings, since the simple MLP project fails to model the complex visual-textual mapping. While updating the LLM parameters allows LLM-RG4 to correctly identify more findings like "atelectasis", the fundamental issue of standard SFT leads to its coarse-grained alignment, which explicitly denies the presence of "pleural effusion" that is clearly visible in the image. All aforementioned issues are alleviated by S2D-ALIGN with comprehensive findings, meanwhile maintaining high descriptive quality. This vividly demonstrates that simply fine-tuning the LLM is insufficient to obtain an anatomically-grounded alignment, where PAG is crucial to empower the model to move beyond coarse pattern matching and perform factually-correct clinical reasoning.

Ablation Studies

Effect of PAG

To validate the design of PAG, we conduct extensive ablation studies by exploring different training paradigms, with results detailed in Table 3. We consider three variants of training paradigms, i.e., training with only the coarse-grained stage ("Single-Stage"), mixing all data for "Joint Training", and training in a "Reversed Order" (S1→S3→S2), along with two baselines ablating the second and third training stages of PAG. It is observed that "Single-Stage" and "Joint Training" result in substantially lower performance, particularly on the F1 score that reflects the factual correctness, indicating that standard SFT (radiograph-report alignment) is insufficient while mixed training struggles to guide the model to learn a coarse-to-fine alignment process. Similar results are seen in "S1 → S2" where fine-grained grounding is removed, showing that injecting key phrases as an auxiliary signal is crucial for the shallow-to-deep learning. Interestingly, by comparing "S1 → S3" and "S1 → S3 → S2", we observe that fine-tuning in a reversed order might harm the performance of directly aligning with key phrases, which underscores a potential impact of PAG similar to that of curriculum learning, with the best results demonstrated by our full model. This suggests that a carefully structured curriculum for domain-specific MLLM, which first establishes a foundational understanding, and then refines it with increasingly granular supervision, is a more effective than simply exposing it to massive data.

Effect of SMA

To investigate the effect of SMA, we replace it with other variants of connector modules, and ablate the SMA design, with results summarized in Table 4. Both the MLP and the MLP with a Q-Former yield drastically lower scores across all metrics, particularly on F1 score, which indicate that conventional connectors struggle to model complex and fine-grained visual-textual mappings for the purpose of RRG. Notably in this comparison, the integration of Q-Former causes further performance degradation, since Q-Former might lead to possible information loss due to feature compression, where this conclusion is consistent with some up-to-date MLLM studies (Lin et al. 2023). While the basic SMA architecture improves the performance owing to increased model parameters (MLP + MSA) or the memory mechanism (SMA w/o Shared Memory), the lack of feature sharing still prevents the model from achieving further anatomically-grounded alignment, where the best results are achieved by our full SMA design, confirming the effectiveness of the feature sharing across different stages of PAG.

Conclusion

In this paper, we addressed the critical challenge of factual correctness in RRG, which is undermined by the coarse-grained alignment in standard SFT-based fine-tuning methods. To this end, we introduced S2D-ALIGN, a novel paradigm centered on our PAG strategy, which employs a shallow-to-deep curriculum to explicitly establish anatomically-grounded alignment. This multi-stage process is effectively unified by our lightweight SMA, designed to integrate multi-granularity guidance and overcome the limitations of simple projection layers. Our comprehensive experiments demonstrate that S2D-ALIGN sets a new state-of-the-art on IU X-RAY and MIMIC-CXR, significantly enhancing the factual reliability and clinical utility of generated reports. Ultimately, this work validates that pursuing anatomically-grounded learning is a pivotal direction for building more trustworthy generative models, where we wish it can serve as a reference work for follow-up studies.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering. In *CVPR*, 6077–6086.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum Learning. In *ICML*, 41–48. New York, NY, USA. ISBN 9781605585161.

Bu, S.; Li, T.; Yang, Y.; and Dai, Z. 2024. Instance-level Expert Knowledge and Aggregate Discriminative Attention for Radiology Report Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14194–14204.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021a. 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)*, 5904–5914. Online.

Chen, Z.; Shen, Y.; Song, Y.; and Wan, X. 2021b. Cross-modal Memory Networks for Radiology Report Generation. In *ACL*, 5904–5914.

Chen, Z.; Song, Y.; Chang, T.; and Wan, X. 2020a. Generating Radiology Reports via Memory-driven Transformer. In *EMNLP*, 1439–1449.

Chen, Z.; Song, Y.; Chang, T.-H.; and Wan, X. 2020b. Generating Radiology Reports via Memory-driven Transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1439–1449. Online.

Chen, Z.; Varma, M.; Delbrouck, J.; Paschali, M.; Blankemeier, L.; Veen, D. V.; Valanarasu, J. M. J.; Youssef, A.; Cohen, J. P.; Reis, E. P.; Tsai, E. B.; Johnston, A.; Olsen, C.; Abraham, T. M.; Gatidis, S.; Chaudhari, A. S.; and Langlotz, C. P. 2024. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. *CoRR*.

Demner-Fushman, D.; Kohli, M. D.; Rosenman, M. B.; Shooshan, S. E.; Rodriguez, L.; Antani, S. K.; Thoma, G. R.; and McDonald, C. J. 2016. Preparing A Collection of Radiology Examinations for Distribution and Retrieval. *J. Am. Medical Informatics Assoc.*, 23(2): 304–310.

Deria, A.; Kumar, K.; Chakraborty, S.; Mahapatra, D.; and Roy, S. 2024. InVERGe: Intelligent Visual Encoder for Bridging Modalities in Report Generation. In *CVPR*, 2028–2038.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. Cite arxiv:2106.09685Comment: Draft V2 includes better baselines, experiments on GLUE, and more on adapter latency.

Huang, Z.; Zhang, X.; and Zhang, S. 2023. KiUT: Knowledge-injected U-Transformer for Radiology Report Generation. In *CVPR*, 19809–19818.

Hyland, S. L.; Bannur, S.; Bouzid, K.; Castro, D. C.; Ranjit, M.; Schwaighofer, A.; Pérez-García, F.; Salvatelli, V.; Srivastav, S.; Thieme, A.; Codella, N.; Lungren, M. P.; Wetscherek, M. T.; Oktay, O.; and Alvarez-Valle, J. 2023. MAIRA-1: A Specialised Large Multimodal Model for Radiology Report Generation. *CoRR*, abs/2311.13668.

Jain, S.; Agrawal, A.; Saporta, A.; Truong, S. Q.; Nguyen Duong, D.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M. P.; Ng, A. Y.; et al. 2021. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports. *arXiv preprint arXiv:2106.14463*.

Jin, H.; Che, H.; Lin, Y.; and Chen, H. 2024. PromptMRG: Diagnosis-driven Prompts for Medical Report Generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2607–2615.

Jing, B.; Wang, Z.; and Xing, E. 2019. Show, Describe and Conclude: On Exploiting the Structure Information of Chest X-ray Reports. In *ACL 2019*, 6570–6580. Florence, Italy.

Jing, B.; Xie, P.; and Xing, E. 2018a. On the Automatic Generation of Medical Imaging Reports. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2577–2586. Melbourne, Australia.

Jing, B.; Xie, P.; and Xing, E. 2018b. On the Automatic Generation of Medical Imaging Reports. In *ACL 2018*, 2577–2586. Melbourne, Australia.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, A De-identified Publicly Available Database of Chest Radiographs with free-text reports. *Scientific data*, 6(1): 317.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597.

Li, Y.; Liang, X.; Hu, Z.; and Xing, E. P. 2018. Hybrid Retrieval-Generation Reinforced Agent for Medical Image Report Generation. In *NeurIPS 2018*, 1537–1547.

Lin, C.-Y. 2004. Rouge: A Package for Automatic Evaluation of Summaries. In *Text summarization branches out*, 74–81.

Lin, Z.; Liu, C.; Zhang, R.; Gao, P.; Qiu, L.; Xiao, H.; Qiu, H.; Lin, C.; Shao, W.; Chen, K.; Han, J.; Huang, S.; Zhang, Y.; He, X.; Li, H.; and Qiao, Y. 2023. SPHINX: The Joint Mixing of Weights, Tasks, and Visual Embeddings for Multi-modal Large Language Models. *CoRR*.

Liu, C.; Tian, Y.; Chen, W.; Song, Y.; and Zhang, Y. 2024. Bootstrapping Large Language Models for Radiology Report Generation. In *AAAI*, 18635–18643.

Liu, C.; Tian, Y.; and Song, Y. 2024. A Systematic Review of Deep Learning-based Research on Radiology Report Generation. arXiv:2311.14199.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2021a. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. arXiv:2106.06963.

Liu, F.; Wu, X.; Ge, S.; Fan, W.; and Zou, Y. 2021b. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 13753–13762.

Liu, F.; Yin, C.; Wu, X.; Ge, S.; Zhang, P.; and Sun, X. 2021c. Contrastive Attention for Automatic Chest X-ray Report Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 269–280. Online.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *NeurIPS*.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning. arXiv:1612.01887.

Nicolson, A.; Dowling, J.; and Koopman, B. 2022. Improving Chest X-Ray Report Generation by Leveraging Warm-Starting. arXiv:2201.09405.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pérez-García, F.; Sharma, H.; Bond-Taylor, S.; Bouzid, K.; Salvatelli, V.; Ilse, M.; Bannur, S.; Castro, D. C.; Schwaighofer, A.; Lungren, M. P.; et al. 2024. Rad-dino: Exploring Scalable Medical Image Encoders Beyond Text Supervision. *arXiv preprint arXiv:2401.10815*.

Qin, H.; and Song, Y. 2022a. Reinforced Cross-modal Alignment for Radiology Report Generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, 448–458. Dublin, Ireland.

Qin, H.; and Song, Y. 2022b. Reinforced Cross-modal Alignment for Radiology Report Generation. In *ACL*, 448–458.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical Sequence Training for Image Captioning. arXiv:1612.00563.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and Explainable Region-guided Radiology Report Generation. In *CVPR*.

Thawkar, O.; Shaker, A.; Mullappilly, S. S.; Cholakkal, H.; Anwer, R. M.; Khan, S.; Laaksonen, J.; and Khan, F. S. 2023. XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models. arXiv:2306.07971.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; Mustafa, B.; Chowdhery, A.; Liu, Y.; Kornblith, S.; Fleet, D. J.; Mansfield, P. A.; Prakash, S.; Wong, R.; Virmani, S.; Semturs, C.; Mahdavi, S. S.; Green, B.; Dominowska, E.; y Arcas, B. A.; Barral, J. K.; Webster, D. R.; Corrado, G. S.; Matias, Y.; Singhal, K.; Florence, P.; Karthikesalingam, A.; and Natarajan, V. 2023. Towards Generalist Biomedical AI. *CoRR*, abs/2307.14334.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and Tell: A Neural Image Caption Generator. arXiv:1411.4555.

Wang, L.; Ning, M.; Lu, D.; Wei, D.; Zheng, Y.; and Chen, J. 2022. An Inclusive Task-aware Framework for Radiology Report Generation. In *MICCAI*.

Wang, Z.; Liu, L.; Wang, L.; and Zhou, L. 2023. R2GenGPT: Radiology Report Generation with Frozen LLMs. *arXiv*.

Wang, Z.; Sun, Y.; Li, Z.; Yang, X.; Chen, F.; and Liao, H. 2025. LLM-RG4: Flexible and Factual Radiology Report Generation Across Diverse Input Contexts. In *AAAI*, 8250–8258.

Yan, B.; Liu, R.; Kuo, D.; Adithan, S.; Reis, E.; Kwak, S.; Venugopal, V.; O'Connell, C.; Saenz, A.; Rajpurkar, P.; and Moor, M. 2023. Style-Aware Radiology Report Generation with RadGraph and Few-Shot Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. MiniGPT-4: Enhancing Vision-language Understanding with Advanced Large Language Models. arXiv:2304.10592.