

M4CXR: Exploring Multitask Potentials of Multimodal Large Language Models for Chest X-Ray Interpretation

Jongwon Park¹, Soobum Kim¹, Byungmu Yoon¹, Jihun Hyun¹, and Kyoyun Choi¹

Abstract—The rapid evolution of artificial intelligence, especially in large language models (LLMs), has significantly impacted various domains, including healthcare. In chest X-ray (CXR) analysis, previous studies have employed LLMs, but with limitations: either underutilizing the LLMs’ capability for multitask learning or lacking clinical accuracy. This article presents M4CXR, a multimodal LLM designed to enhance CXR interpretation. The model is trained on a visual instruction-following dataset that integrates various task-specific datasets in a conversational format. As a result, the model supports multiple tasks such as medical report generation (MRG), visual grounding, and visual question answering (VQA). M4CXR achieves state-of-the-art clinical accuracy in MRG by employing a chain-of-thought (CoT) prompting strategy, in which it identifies findings in CXR images and subsequently generates corresponding reports. The model is adaptable to various MRG scenarios depending on the available inputs, such as single-image, multiimage, and multistudy contexts. In addition to MRG, M4CXR performs visual grounding at a level comparable to specialized models and demonstrates outstanding performance in VQA. Both quantitative and qualitative assessments reveal M4CXR’s versatility in MRG, visual grounding, and VQA, while consistently maintaining clinical accuracy.

Index Terms—Chest X-ray (CXR), medical imaging, medical report generation (MRG), multitask learning, multimodal large language model (LLM), visual grounding, visual question answering (VQA).

I. INTRODUCTION

RECENT advancements in artificial intelligence, particularly large language models (LLMs), have led to their widespread application across various fields, including healthcare. Numerous studies are exploring diverse methods of application for improving healthcare outcomes, such as personalized treatment plans, clinical decision support systems, and enhancing medical education [1].

Within the medical domain, this article focuses on chest X-ray (CXR) interpretation. Although many studies have

Received 6 January 2025; revised 13 May 2025; accepted 29 June 2025. Date of publication 1 August 2025; date of current version 9 October 2025. This work was supported by the Technology Innovation Program (RS-2025-02221011, Development of Medical-Specialized Multimodal Hyperscale Generative AI Technology for Global Integration) funded by the Ministry of Trade Industry & Energy (MOTIE, South Korea). (Corresponding author: Kyoyun Choi.)

The authors are with DEEPNOID Inc., Seoul 08377, South Korea (e-mail: jgpark@deepnoid.com; soobumk@deepnoid.com; bmyoon@deepnoid.com; jhyun@deepnoid.com; kychoi@deepnoid.com).

This article has supplementary downloadable material available at <https://doi.org/10.1109/TNNLS.2025.3587687>, provided by the authors.

Digital Object Identifier 10.1109/TNNLS.2025.3587687

investigated LLMs for CXR, most encounter one of two issues: they either underutilize the capabilities of LLMs or struggle with ensuring clinical accuracy. Despite LLMs’ ability to perform various tasks through conversational interaction, many studies focus on a single task, typically medical report generation (MRG) [2], [3], [4]. While some studies effectively leverage LLMs to comprehend questions and generate appropriate responses [5], [6], [7], [8], the clinical accuracy of these conversational outcomes, including generated reports, falls short of expectations.

In this work, we propose M4CXR, an LLM for CXR adept at handling four “Multi” aspects: Multimodal, Multitask, Multiimage input, and Multiturn chatting. We train the model on a dataset constructed by integrating various task-specific datasets, enabling M4CXR to excel in MRG, image understanding, and visual question answering (VQA) tasks, as depicted in Fig. 1. In MRG, M4CXR improves clinical accuracy by leveraging LLMs’ reasoning abilities through a chain-of-thought (CoT) [9] reasoning, where the model identifies findings in the image and generates reports based on these results. Additionally, we conduct experiments across various scenarios with inputs extending beyond a single image, including CXR images from different views and prior patient studies. Besides MRG, M4CXR demonstrates a strong capability for multitask learning with successful application in visual grounding and VQA, showcasing its adaptability in diverse clinical contexts.

In summary, our main contributions are as follows.

- 1) We propose M4CXR, an MLLM designed for CXR interpretation, capable of handling multiple tasks. To enable multitask learning, we assemble a visual instruction-following dataset from diverse CXR tasks.
- 2) By adopting a novel CoT reasoning process, M4CXR achieves state-of-the-art clinical accuracy in CXR report generation.
- 3) M4CXR effectively utilizes multiple images and reports, allowing for its applicability across different scenarios.
- 4) Beyond MRG, M4CXR demonstrates remarkable performance in visual grounding and VQA.

II. RELATED WORKS

A. Multimodal LLMs

As research on LLMs progresses, their advantages continue to emerge. It is widely known that LLMs are capable of

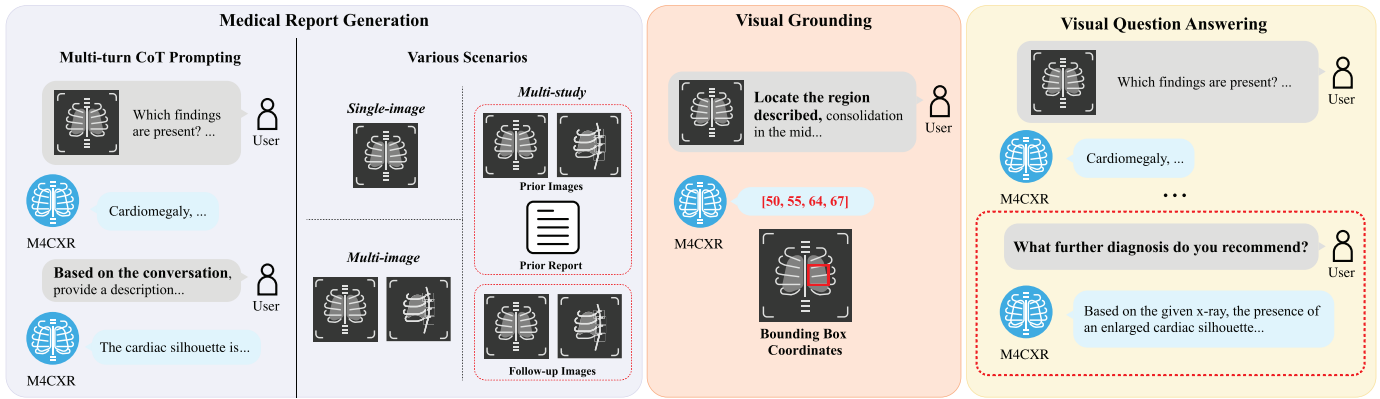


Fig. 1. Overview of M4CXR’s capability for multitask learning. Facilitated by CoT prompting in MRG, M4CXR produces clinically accurate reports and adapts to various scenarios. Additionally, M4CXR can ground the locations described in the report or answer questions based on CXR images.

multitask learning [10], and they can also reason through multiturn chats using CoT prompting [9]. The rise of multimodal LLMs, which utilize the long context length of LLMs to process visual content, further amplifies these advantages. MLLMs can handle tasks that require comprehension of spatial regions, and visual instruction tuning [11] allows them to answer free-form questions about images.

For MLLMs, bridging the embedding spaces of vision and language is crucial. A common approach is to freeze the pre-trained vision model and LLM, and train only the bridging module [12], [13], also known as the projector. The projector’s structure varies, from a linear layer to Q-Former [14], C-Abstractor [15], and more. Drawing insights from these works, we employ a two-phase training strategy: pre-training the projector first, and then fine-tuning the entire model with visual instruction tuning.

B. CXR Report Generation

CXR interpretation via deep learning has been extensively studied over time [16], [17], [18], [19]. Since the advent of LLMs, strategies such as freezing pre-trained models [5] or visual instruction tuning [7] to train CXR report generation models have been equivalently implemented. However, a significant drawback of these models is the lack of clinical accuracy in the generated reports. Efforts to address these shortcomings include adopting a separate disease classification module [2], [20], increasing the visual encoder’s resolution [4], or incorporating additional input such as anatomy and clinical prompts [21] or medical context information [22]. In this work, we take a novel yet simple approach: leveraging the reasoning abilities of LLMs.

C. Multitask Learning in CXR Interpretation

Confining the use of LLMs solely to report generation underutilizes their potential. MLLMs that capitalize on the characteristics of LLMs to enable conversations based on CXR images include XrayGPT [5], RaDialog [6], and LLM-CXR [7]. CheXagent [8] is a CXR foundation model trained on various tasks, similar to our approach. Yet, the clinical accuracy of these models’ responses remains suboptimal.

Medical foundation models such as Med-Gemini [23] and MedPaLM-M [24] support CXR interpretation with conversational capabilities, showing satisfactory clinical accuracy. However, these models are not trained on spatial cognition tasks like visual grounding, which involves identifying regions in an image corresponding to a phrase in the text. For CXR, the MS-CXR benchmark dataset [25] was released for phrase grounding, and MedRPG [26] specializes in identifying bounding boxes for phrases in CXR images. The latest concurrent work introduced MAIRA-2 [27], an MRG model for CXR capable of grounding but limited to grounded report generation. In contrast, M4CXR handles multiple tasks such as MRG, VQA, and visual grounding, while maintaining clinical accuracy.

III. METHODS

A. M4CXR

1) *Architecture*: Fig. 2(a) illustrates the overall architecture of M4CXR. Following LLaVA [11], M4CXR includes a vision encoder, a projector, and an LLM, denoted as \mathcal{M}_V , \mathcal{M}_P , and \mathcal{M}_L , respectively. n CXR images x_1, \dots, x_n and a text prompt with n image slots, inserted in appropriate positions based on the context, are given. The prompt is mapped into a sequence of language token embeddings X^l . The vision encoder processes each image x_i to extract its visual feature f_i^v , which is subsequently transformed by the projector into a sequence of visual token embeddings X_i^v , ensuring dimensional alignment with X^l

$$X_i^v = \mathcal{M}_P(f_i^v) = \mathcal{M}_P(\mathcal{M}_V(x_i)), \quad \text{for } i = 1, \dots, n. \quad (1)$$

Taking as input the combined sequence of X^l and X_i^v s, with X_i^v s placed in the image slots, the LLM generates output Y

$$Y = \mathcal{M}_L(X_1^v, \dots, X_n^v, X^l). \quad (2)$$

The output of the MLLM $Y = \{y_t\}_{t=1}^T$, consisting of T language tokens, is generated autoregressively

$$p(Y|X_1^v, \dots, X_n^v, X^l) = \prod_{t=1}^T p(y_t|X_1^v, \dots, X_n^v, X^l, y_{<t}). \quad (3)$$

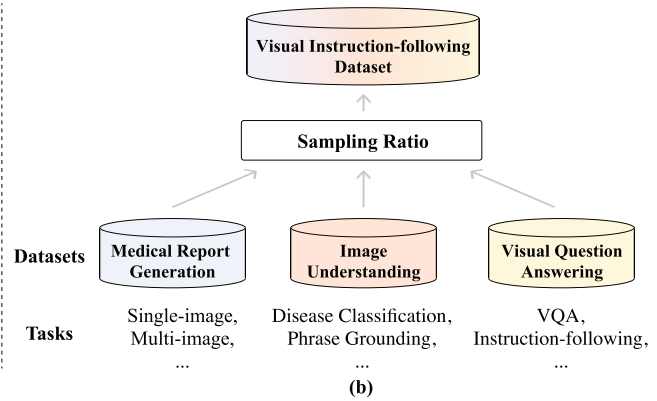
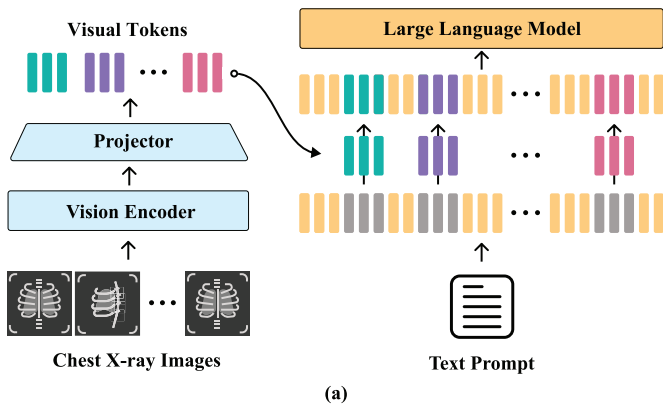


Fig. 2. (a) Architecture of M4CXR. Utilizing the LLaVA framework, it allows visual tokens from each image to be inserted at designated positions among the text tokens. (b) Schema for constructing a CXR visual instruction-following dataset. Diverse tasks of three types are combined with appropriate sampling ratios.

2) *Bounding Box Representation*: To enable visual grounding, MLLMs need to represent spatial information as text tokens. We use bounding box coordinates enclosed in square brackets without any additional special tokens, following the approach in [28]. The coordinates $[x_1, y_1, x_2, y_2]$ represent the top-left (x_1, y_1) and bottom-right (x_2, y_2) points of the bounding box on the image. As input images are preprocessed to a uniform size by the vision encoder, the bounding box coordinates are also normalized [29] to integer values between 0 and 100.

B. Multiturn CoT Prompting

We derive insights from PromptMRG [20] to enhance the clinical accuracy of MRG. It classifies observable lesions from a CXR image and then uses the result as input prompts to generate reports. We follow a similar two-step process, first identifying key observations and then generating reports. The difference from PromptMRG is that, in our work, a single MLLM performs both classification and report generation sequentially, enabling end-to-end learning.

The gist is to divide the input prompt into two questions, creating multiturn conversational data. In the first question, we present a list of potential observations and prompt the model to identify those visible in the CXR image. Then, we ask the model to generate a report based on the prior conversation. This approach resembles the reasoning process of a human radiologist, who first identifies lesions visible in the image and then generates a diagnostic report. It can be seen as a variant of the commonly used CoT [9] prompting, hence we call it CoT MRG. Fig. 3 shows the multiturn CoT prompting process. To validate the effectiveness of CoT MRG, we conduct a comparative experiment with classification and MRG as separate tasks.

C. Integrating CXR Interpretation Tasks

We focus on the following features of an MLLM: the capability for reasoning through multiturn chat, the ability to utilize multiple images due to its long context length, and the flexibility for multitask learning. To exploit these features,

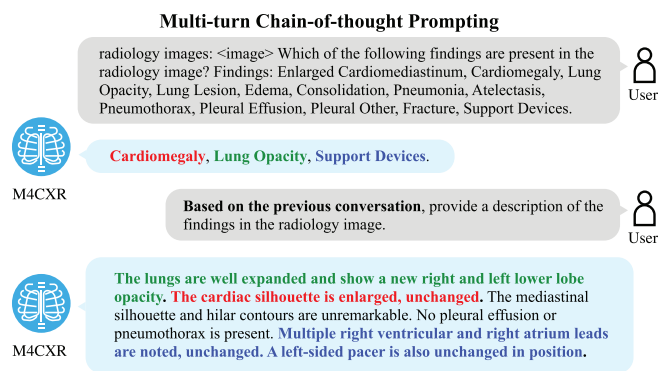


Fig. 3. Example of multiturn CoT prompting. M4CXR first identifies findings in the CXR image, then generates a report. Findings and corresponding sentences are color-matched for readability.

we construct a CXR visual instruction-following dataset, as schematized in Fig. 2(b). We define a task set across three task types: MRG, image understanding, and VQA, and transform the corresponding datasets into conversational data. Detailed descriptions for all CXR interpretation tasks considered in this work and the corresponding conversation templates are listed in Table I and in Appendix, respectively.

1) *Report Generation: Various Scenarios*: A single radiologic study can include multiple CXR images taken from different views: posterior-anterior (PA), anterior-posterior (AP), or lateral. Additionally, when the same patient undergoes follow-up studies, information from prior studies can also be utilized in the radiologic interpretation.

In contrast to previous studies that generate reports from a single image [3], [4], [17], [20], we consider MRG tasks in three different scenarios based on the available inputs. The *single-image* scenario incorporates one image from a single study. The *multiimage* scenario accesses multiple images from different views within a single study. The *multistudy* scenario accepts images and the corresponding report from prior studies as inputs, along with follow-up study images. These scenarios have different conversation templates. In this work, since we distinguish tasks by templates, they are treated as separate

TABLE I
TASK DESCRIPTION

Task Type	Task	Description
Medical Report Generation	<i>Single-image</i>	Generates the FINDINGS section of a report from a single CXR image.
	<i>Multi-image</i>	Generates the FINDINGS section of a report from one or more images, including images from different views within the same study.
	<i>Multi-study</i>	Generates the FINDINGS section of a report from one or more studies, combining current and previous studies from a patient.
Image Understanding	Disease Classification	Identifies diseases within a single image based on the labels covered by each dataset.
	Finding Grounding	Provides the bounding box coordinates for a given finding, if detected.
	Grounded Finding	Identifies the finding corresponding to a given bounding box.
	Abnormality Detection	Identifies and localizes abnormal regions by providing the corresponding bounding box coordinates.
	Multi Finding Grounding	Identifies the presence of candidate findings and provides the bounding box coordinates for each detected finding.
	Organ Grounding	Provides the bounding box coordinates for a given organ.
	Grounded Organ	Identifies the organ corresponding to a given bounding box.
	Grounded Phrase Generation	Generates a radiology report phrase for the region corresponding to a given bounding box.
	Phrase Grounding	Provides the bounding box coordinates for the region described by a given phrase.
	Anatomical Region Grounding	Provides the bounding box coordinates for a given anatomical region.
Visual Question Answering	Grounded Anatomical Region	Identifies the anatomical region corresponding to a given bounding box.
	Visual Question Answering	Responds to various questions about the content of a radiology image.
	Difference Visual Question Answering	Compares past (reference) and current (main) images to answer questions about the differences between them.
	Visual Instruction-following	Generates responses by following specific instructions, facilitating clear communication in multiple interactions.

tasks. Through these different scenario tasks, we expect the model to understand and fully utilize the given information, enabling it to generate more accurate reports when additional inputs are available.

2) *Image Understanding*: Disease classification is one of the core tasks in CXR image understanding. We use class label data from various classification datasets. Since each dataset has its own set of disease labels, we specify the relevant labels for each question in the prompt. The model is prompted to identify findings in the image from a list of diseases.

Our task set also includes fine-grained tasks requiring detailed analysis of specific CXR regions. These tasks involve grounding and identification of findings, phrases, organs, and anatomical structures, as well as abnormality detection. We convert bounding box labels and clinical text reports into instruction-following data to help the model link CXR images with spatial information for detailed analysis.

3) *Visual Question Answering*: To train M4CXR, we convert all the data into an instruction-following format using predefined templates. However, training the LLM solely on templates presents a risk of overfitting: the model might lose its inherent conversational abilities and respond only in a fixed format. Therefore, we also capitalize on VQA datasets, which are already in a conversational format, expecting the model to freely respond to a diverse range of questions.

IV. EXPERIMENTS

A. Training Datasets

We collect and integrate various datasets according to the tasks described in Section III.

1) *Report Generation*: MIMIC-CXR [31] encompasses a diverse collection of CXR images along with detailed radiology reports. These reports are extensively annotated, facilitating advanced medical image analysis. For MRG tasks, our focus is on generating the FINDINGS section, which offers an in-depth description of significant observations identified in the CXR images.

We employ CheXbert [32] to extract observation labels from the FINDINGS section. CheXbert outputs four-class (positive, negative, uncertain, blank) classification results for 14 predefined observation labels. For binary classification, all non-positive classes are treated as negative. Its label set of observations is provided as the candidate findings in the first question of CoT prompting.

For the three different MRG scenario tasks, image-report pairs are organized as follows. For *single image*, every image is used as a data instance. In *multiimage*, images from different views that share the same study ID are gathered to compose study-level data. *Multistudy* combines two consecutive studies of a patient, with the studies arranged in chronological order.

2) *Image Understanding*: BRAX [33] and CheXpert [30] datasets are used for disease classification. We incorporate

TABLE II
TRAINING HYPERPARAMETERS

Configuration	Pre-training	Instruction tuning
Training module	Projector	Vision Encoder, Projector, LLM
Training steps	2k	40k
Max steps	10k	100k
Warmup steps	500	
Global batch size	256	64
Optimizer	AdamW	
Optimizer hyperparameter	$\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 1e^{-8}$	
Learning rate	$3e-4$	$1e-4$
Learning rate schedule	Cosine decay	
Weight decay	$1e-2$	$1e-4$
Gradient clipping	1.0	

datasets that contain bounding boxes along with class labels (VinDr-CXR [34], ChestX-ray14 [35], ChestX-Det10 [36], JSRT [37], SIIM [38], RSNA [39], COVID-19 Radiography [40], COVID-QU-Ex [40], [41], [42], QaTa-COV19 [43]) for various fine-grained image understanding tasks in addition to disease classification. MS-CXR [25] includes image-sentence pairs of bounding boxes and corresponding phrases for images and reports from MIMIC-CXR. ImaGenome [44], which is also derived from MIMIC-CXR, annotates text in the report with bounding boxes aligned with 29 anatomical regions.

3) *Visual Question Answering*: MIMIC-CXR-VQA [45] and MIMIC-Diff-VQA [46] datasets are designed to address a wide range of questions based on MIMIC-CXR images. These datasets encompass diverse and comprehensive question-answer pairs, enabling models to effectively handle various CXR-related queries. RaDialog [6] is a visual instruction-following dataset designed to facilitate tasks that require structured dialogs. It guides models in understanding and responding to instructions based on CXR images.

B. Test Datasets

The test set used for evaluating all MRG tasks is from MIMIC-CXR. The official test split includes 2461 images in the frontal (PA, AP) view and 1397 in other views, totaling 3858 images. Each study corresponding to the 2461 frontal images is considered a separate data instance in *multiimage* and *multistudy*, with the prior study included for the latter if available. As a result, the test set size for *multiimage* and *multistudy* is 2461, while *single-image* considers two test set sizes: 2461 and 3858.

For phrase grounding evaluation, the test set of MS-CXR is used. Each query phrase in MS-CXR corresponds to a single bounding box. To ensure a rigorous evaluation of the task, we exclude all MS-CXR images from the training set of other datasets across all tasks. Following the data split in [26], the test set has 167 images.

We employ the test set of MIMIC-CXR-VQA for evaluating medical VQA. Of the 13793 total samples, we exclude 2484 samples that lack answers, resulting in 11309 samples for assessment. Additionally, we incorporate SLAKE [47], which is not included in the training data. From its test set of 2094 samples, we use only CXR images with close-ended questions in English, retaining 114 samples.

C. Evaluation Metrics

1) *Report Generation*: The generated reports are assessed using natural language generation (NLG) and clinical metrics. We use the traditional BLEU [48] and ROUGE-L [49] for NLG metrics.

Regarding clinical accuracy, we calculate $F1$ -scores from CheXbert classification results. While example-based $F1$ (eF1) calculates and averages $F1$ for each example [50], macro-averaged $F1$ (MF1) computes $F1$ per class and averages across classes, and micro-averaged $F1$ (mF1) aggregates all class contributions before calculating $F1$. $F1$ -14 indicates the $F1$ -score calculated over all 14 CheXbert labels, while $F1$ -5 is computed for only 5 of those labels (cardiomegaly, edema, consolidation, atelectasis, pleural effusion).

RadGraph-F1 [51] is another clinical accuracy metric. RadGraph [52] extracts a graph composed of clinical entities and relations from a radiology report. RadGraph-F1 is the average of $F1$ -scores for the graph’s entities and relations.

2) *Medical Phrase Grounding*: To evaluate the performance of phrase grounding, we calculate the intersection over union (IoU) between the ground-truth and predicted bounding boxes. We use the mean IoU (mIoU) averaged over all data as the evaluation metric. Additionally, accuracy is determined by considering predicted boxes with an IoU of 0.5 or higher as correct predictions.

3) *Visual Question Answering*: Accuracy, recall, and BLEU-1 are used to assess the VQA performance. Exact matches between predictions and ground truth are counted to calculate the accuracy. To consider partial matches for open-ended responses, recall is calculated by measuring the proportion of ground-truth words present in the generated sequences. For MIMIC-CXR-VQA, which includes open-ended questions, BLEU-1 scores are also calculated for each test sample and then averaged.

D. Implementation Details

We use a randomly initialized C-Abstractor as the projector to efficiently handle multiple radiology images and reduce image tokens. The LLM and vision encoder are pre-trained models: Mistral-7B-Instruct-v0.2 [53] and Rad-DINO [54], respectively. The training involves two stages. First, we pre-train the projector while keeping the LLM and vision encoder frozen, using only CXR images and reports without instruction prompts. Subsequently, the vision encoder, projector, and LLM are trained together for visual instruction tuning. We apply LoRA [55] for LLM training to reduce computational cost. Each instruction-following data input includes images and instruction texts, and the model is trained to predict the corresponding responses using cross-entropy loss.

TABLE III

SINGLE-IMAGE MRG PERFORMANCE ON THE MIMIC-CXR TEST SET. THE BEST PERFORMANCES ARE HIGHLIGHTED IN BOLD. THE NUMBERS FOR CHEXAGENT* WERE OBTAINED FROM [4]. MODELS WITH † REPORT CLASSIFICATION F1-SCORES USING A DIFFERENT LABELER, CHEXPRT [30]

Model	CheXbert					Rad-Graph	BLEU		ROUGE	Test set
	mF1-14	mF1-5	MF1-14	MF1-5	eF1-14	F1	-1	-4	-L	size
LLM-CXR†	36.0	-	21.1	-	-	-	9.2	1.5	16.2	3,530
RaDialog	-	-	39.4	-	-	-	34.6	9.5	27.1	-
METransformer†	-	-	-	-	31.1	-	38.6	12.4	29.1	3,269
DCL†	-	-	-	-	37.3	-	-	10.9	28.4	-
PromptMRG	-	-	38.1	-	47.6	-	39.8	11.2	26.8	3,858
LM-RRG	-	-	-	-	48.4	-	-	12.2	29.6	-
Med-PaLM M 84B	53.6	57.9	39.8	51.6	-	26.7	32.3	11.3	27.5	4,834
CheXagent*	39.3	41.2	24.7	34.5	-	-	16.9	4.7	21.5	2,461
MAIRA-1	55.7	56.0	38.6	47.7	-	24.3	39.2	14.2	28.9	2,461
LLaVA-Rad	57.3	57.4	39.5	47.7	-	-	38.1	15.4	30.6	2,461
M4CXR	58.1	61.6	38.8	49.5	50.2	21.7	33.3	10.2	28.4	3,858
	60.6	61.8	40.0	49.5	53.6	21.8	33.9	10.3	28.5	2,461

The detailed hyperparameters for training are summarized in Table II. C-Abstractor is configured with a depth of 3, an MLP depth of 2, and a hidden size of 1024, to be employed as the projector. The number of visual tokens is set to 361, corresponding to a 19×19 grid. Accordingly, the output of the vision encoder, which is a sequence of 1369 (37×37) tokens processed by Rad-DINO, is compressed to a length of 361. This reduction in the number of tokens, by approximately one-fourth, enables the LLM to handle a greater number of images more efficiently. In the training of the LLM, we configure the LoRA hyperparameters as follows: rank is set to 8, alpha to 32, and dropout to 0.05. Training our 7.37 billion parameter model on two H100 GPUs using bfloat16 with automatic mixed precision and flash-attention v2 [56] takes approximately 24 h for the first stage and 80 h for the second stage. For evaluation, sentences are generated with greedy search decoding, except for the qualitative evaluation of VQA: to generate more natural responses, multinomial sampling is applied. The average inference speed is approximately 1.24 s/image for *single-image* CoT MRG. The source code is available at: <https://github.com/deepnoid-ai/M4CXR-TNNLS>.

V. RESULTS

A. Medical Report Generation

1) *Single Image*: To evaluate *single-image* MRG performance, we compared our model with state-of-the-art MRG models, including LLM-CXR, RaDialog, METransformer [57], DCL [58], PromptMRG, LM-RRG [2], CheXagent, MAIRA-1 [3], Med-PaLM M, and LLaVA-Rad [4]. The results are summarized in Table III. For direct comparison with other models, we evaluated M4CXR using two distinct test sets: one with 2461 frontal images and another with 3858 images of all views. Including lateral views decreased clinical accuracy, suggesting the complexity of recognizing observations from lateral images.

Among the models evaluated on only frontal images, M4CXR attained the highest CheXbert clinical accuracy, with

mF1-14 and MF1-14 scores of 60.6 and 40.0, respectively. In the evaluation that included all views, our model outperformed PromptMRG with an eF1-14 of 50.2. Although Med-PaLM M used a different all-view test set, which limits direct comparison, our model showed a significant advantage in mF1, with similar or slightly lower MF1s.

However, M4CXR did not achieve the best scores in both RadGraph-F1 and NLG metrics. To improve clinical accuracy, CoT prompting provides candidate findings in the first question before generating reports. This approach may have led the model to use terms from the findings list rather than words that exactly match those in the ground-truth report, which could be a contributing factor to the observed lower NLG scores. RadGraph-F1, calculated by extracting entities and relations from the report, was likely reduced due to this mismatch.

2) *Various Scenarios*: Table IV shows the evaluation results across various input scenarios. *Multimage* improved clinical accuracy compared to *single-image*, with CheXbert mF1-14 and MF1-14 scores of 61.1 and 41.0, respectively. Providing the prior study as additional input improved MF1, though mF1 slightly decreased. These results demonstrate that our model can effectively utilize available inputs to generate medical reports in various scenarios.

Although test conditions differ, we compared our model's performance with two baseline models. Tu et al. [24] reported the zero-shot generalization results of Med-PaLM M in a two-view setting. Since it was trained only on *single image*, its effectiveness diminished, unlike M4CXR leveraging additional images for enhanced outcomes. MAIRA-2 accepts a prior study and multiple images as input, akin to our *multistudy* scenario. Its performance matches ours closely: MAIRA-2 achieved a higher MF1-14 of 42.7, whereas M4CXR surpassed it with an mF1-14 of 60.7.

B. Medical Phrase Grounding

Table V compares the medical phrase grounding results of TransVG [59], MedRPG, MAIRA-2, and M4CXR, on

TABLE IV

EVALUATION RESULTS OF MRG PERFORMANCE IN VARIOUS SCENARIOS

Model	MRG Scenario	CheXbert	
		mF1-14	MF1-14
Med-PaLM M 84B MAIRA-2 7B	<i>Multi-image</i>	50.5	37.8
	<i>Multi-study</i>	58.5	42.7
M4CXR	<i>Single-image</i>	60.6	40.0
	<i>Multi-image</i>	61.1	41.0
	<i>Multi-study</i>	60.7	42.0

TABLE V

REFERRING EXPRESSION GROUNDING RESULTS. NOTE THAT MAIRA-2 7B WAS EVALUATED ON A DIFFERENT TEST SPLIT

Model	Accuracy	mIOU
TransVG	65.9	58.9
MedRPG	69.9	59.4
MAIRA-2 7B	-	57.8
M4CXR	68.3	57.9

TABLE VI
MEDICAL VQA PERFORMANCE

Model	MIMIC-CXR-VQA			SLAKE	
	Acc	Recall	BLEU-1	Acc	Recall
RaDialog	0.0	43.0	5.0	0.0	45.6
RadFM	11.2	36.7	13.8	68.4	69.7
CheXagent	59.0	72.8	62.5	71.1	73.2
M4CXR	62.3	70.3	66.4	85.1	86.0

the MS-CXR test set. While MedRPG achieved the highest accuracy and mIoU, MedRPG and TransVG are specialized for phrase grounding and cannot perform other tasks. Given this limitation, M4CXR, which can handle multiple tasks, shows competitive performance with an accuracy of 68.3 and mIoU of 57.9. MAIRA-2, utilizing a private dataset for grounded report generation, reported a similar mIoU to ours.

C. Visual Question Answering

Table VI presents the results of medical VQA evaluation. We conducted our own evaluations on open-source medical MLLMs: LLaVA version of RaDialog [60], RadFM [61], and CheXagent. M4CXR outperformed the other models, except for the recall on MIMIC-CXR-VQA. CheXagent included the dataset in its training, likely explaining its high recall of 72.8. The lower accuracy on MIMIC-CXR-VQA, compared to SLAKE, can be explained by the differences of the two test sets: while the closed-ended questions in SLAKE are limited, the questions in MIMIC-CXR-VQA have greater variety. Consequently, the generated answers also exhibit greater diversity, making it more challenging to generate an accurate response. Moreover, the higher BLEU-1 score compared to accuracy suggests the presence of partially correct answers.

D. Exploration of Sampling Ratios

In multitask learning, the data sampling strategy plays a crucial role, with numerous such strategies being applicable. We experiment with various sampling strategies, as shown in Table VII. The tasks are grouped into three task types, and a dataset used within a single task is referred to as a task-dataset. The per-task-dataset (D1) strategy samples uniformly across all task-datasets. The per-size (D2) strategy samples proportionally based on the size of each task-dataset. After setting equal proportions for each task type (per-task-type), within each task type, D3 samples per-task-dataset, while D4 samples per-size.

In D1, which follows the per-task-dataset strategy, MRG accounts for only 3 out of 50 total task-datasets. This very low sampling ratio for MRG explains its low MRG performance. In contrast, in the per-task-type settings of D3 and D4, one-third of the overall tasks are allocated to MRG, leading to improved MRG performance compared to D1. For grounding performance, D2 and D4 showed inferior results, likely because the training split of MS-CXR, which is the dataset used for evaluation, was barely sampled when using the per-size setting. As for VQA, since all datasets are large in size, per-size sampling (D2) exhibited a high BLEU-1 score. Similar to MRG, VQA also constitutes a small proportion of the total task-dataset (4 out of 50), resulting in lower performance for D1, while D3 and D4 achieved higher scores as per-task-type allocates one-third to VQA.

We aimed to find the appropriate sampling ratios where M4CXR remains clinically accurate, supports visual grounding, and retains the conversational abilities of the LLM. To improve MRG performance, we reduced the ratio allocated to VQA and increased the proportion of MRG. For image understanding, we began with the per-size ratio and made further adjustments to enhance grounding performance. Additionally, to improve the instruction-following capabilities of the model, we increased the ratio of RaDialog compared to other VQA datasets. While this adjustment led to a decrease in quantitative VQA performance metrics, it contributed to preserving the LLM’s ability to understand and respond to instructions in conversation. As a result, the empirical ratio for multitask learning was determined to be approximately 54%, 35%, and 11% for MRG, image understanding, and VQA, respectively.

E. Ablation Study

We examined the effects of multiturn CoT prompting and the combination of task types through comparative experiments (Table VIII). Generating reports without CoT prompting in M4CXR led to a significant drop in clinical accuracy. Exp1, trained on single-turn MRG without the first question of CoT prompts, yielded similar results. Exp2 involved training on single-turn MRG and disease classification as separate tasks. Testing Exp2 with CoT prompts improved clinical accuracy, but NLG metrics fell significantly, with the generated text being merely a list of identified observations. The experiments reveal the effectiveness of multiturn CoT prompting in MRG’s training and testing phases.

TABLE VII
PERFORMANCE COMPARISON BETWEEN DIFFERENT DATA SAMPLING STRATEGIES

	Sampling Strategy	CheXbert		NLG		Grounding	VQA
		mF1-14	MF1-14	BLEU-4	ROUGE-L	mIoU	BLEU-1
M4CXR	empirical ratio	60.6	40.0	10.3	28.5	57.9	66.4
D1	per-task-dataset	52.7	32.1	9.1	26.8	58.5	72.8
D2	per-size	58.5	36.1	9.0	27.2	47.9	77.4
D3	per-task-type & per-task-dataset	59.0	37.1	10.0	28.2	57.9	78.3
D4	per-task-type & per-size	59.1	37.7	9.1	27.3	46.4	78.0

TABLE VIII

RESULTS OF ABLATION STUDIES. “IMGUND” STANDS FOR “IMAGE UNDERSTANDING.” \checkmark^s AND $\checkmark^{s,c}$ INDICATE SINGLE-TURN MRG WITHOUT CoT PROMPTING AND THE TRAINING OF SINGLE-TURN MRG AND DISEASE CLASSIFICATION AS SEPARATE TASKS, RESPECTIVELY

	Training			Test	CheXbert		NLG		Grounding	VQA
	MRG	ImgUnd	VQA		CoT	mF1-14	MF1-14	BLEU-4		
M4CXR	\checkmark	\checkmark	\checkmark	\checkmark	60.6	40.0	10.3	28.5	57.9	66.4
Exp1	\checkmark^s	\checkmark	\checkmark	\times	50.0	33.1	9.3	27.7	56.1	66.1
Exp2	$\checkmark^{s,c}$	\checkmark	\checkmark	\times	58.5	37.5	0.6	10.9	56.3	66.6
Exp3	\checkmark		\checkmark	\checkmark	60.6	38.2	10.1	28.2	-	67.9
Exp4	\checkmark	\checkmark		\checkmark	59.3	38.3	10.1	28.0	57.9	36.1

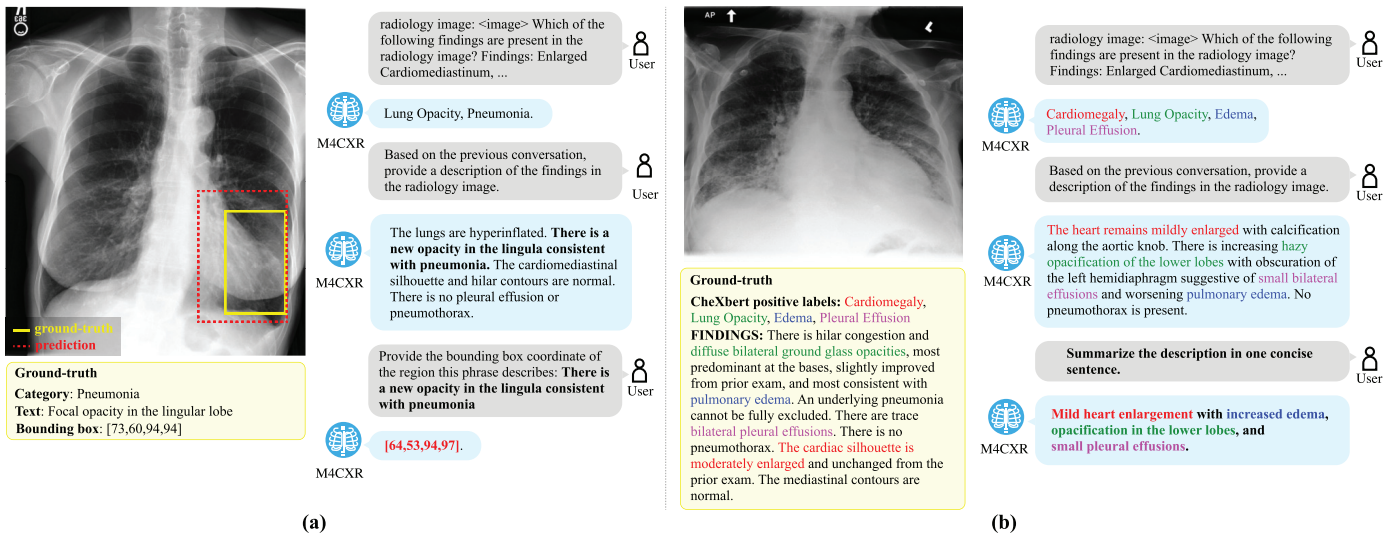


Fig. 4. Examples of M4CXR’s performance in (a) visual grounding and (b) VQA. The images are selected from the test splits of MS-CXR and MIMIC-CXR, respectively.

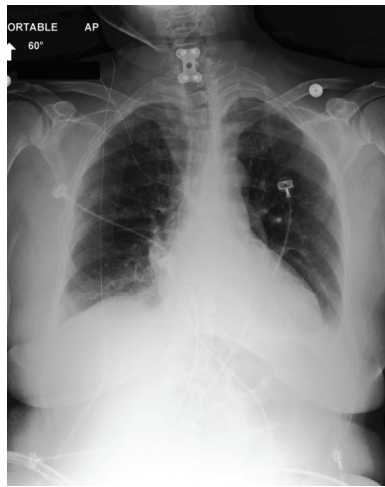
Exp3 and Exp4 excluded image understanding and VQA tasks, respectively. In Exp3, while MRG performance was on par and VQA performance improved due to the increased proportion of VQA data, the absence of image understanding data precluded visual grounding. In Exp4, the BLEU-1 score for VQA dropped significantly from 66.4 to 36.1, suggesting the necessity of VQA datasets in understanding and answering free-form questions. Moreover, the slight decrease in Exp4’s clinical accuracy compared to M4CXR

implies the contribution of VQA datasets to MRG performance improvement.

F. Qualitative Analysis

Figs. 4–6 collectively illustrate examples of dialogs between a user and M4CXR. Additional examples are provided in the Supplementary Material.

In Fig. 4(a), after the multiturn CoT prompting to generate a report, the user asks a third question to locate the region of a



Ground-Truth
CheXbert positive labels: Atelectasis, Pleural Effusion.
FINDINGS: In comparison with the study of ____, there is slightly less opacification at the left base. This is most consistent with atelectasis and effusion, though in the appropriate clinical setting, supervening pneumonia would have to be considered. Specifically, no evidence of pneumothorax. No acute focal pneumonia.

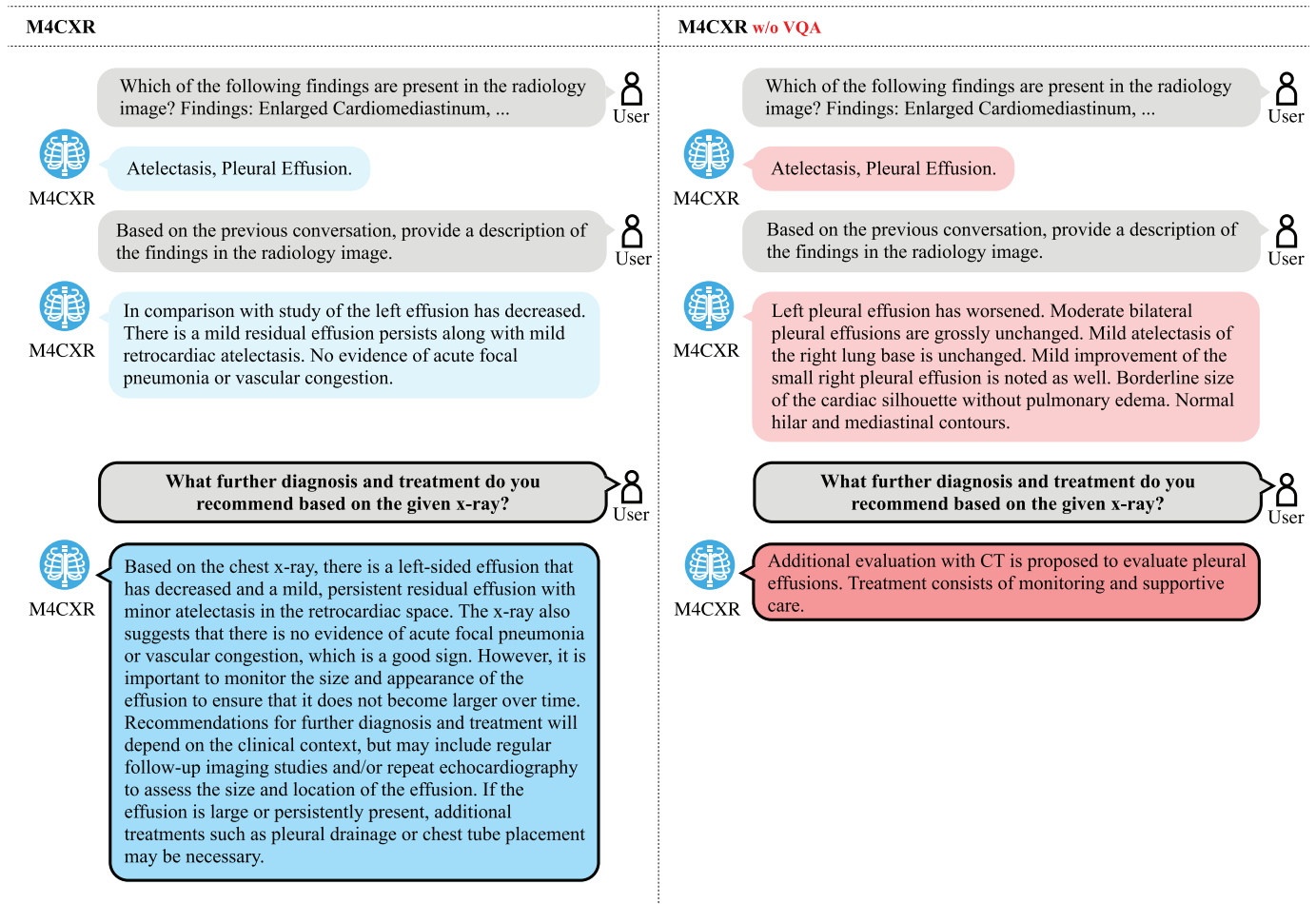


Fig. 5. Examples of medical treatment recommendations. The left shows the results from M4CXR, while the right shows the results when trained without VQA datasets.

sentence in the report. The model provides the coordinates of a bounding box, depicted as a red dashed box. The prediction covers all of the ground-truth, represented by the yellow solid box, possessing sufficient explanatory power to indicate what the corresponding sentence represents.

Fig. 4(b) illustrates M4CXR’s ability to write a clear and concise summary from the generated report. On the left, the

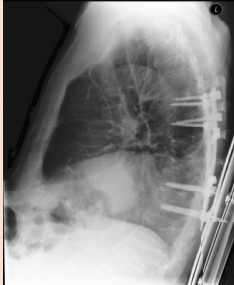
ground-truth report is provided for comparison. For enhanced visual clarity, findings and corresponding sentences are color-matched. M4CXR accurately identifies all the observations in the first question and subsequently generates a comprehensive report without omitting any detail. In response to the third question, which asks for a one-sentence summary, the model offers a concise statement that includes every noted finding.

Ground Truth

CheXbert positive labels: Atelectasis, pleural effusion, support devices.

FINDINGS: A left central venous catheter is seen terminating in the lower SVC. Again seen is elevation of the right hemidiaphragm and small pleural effusion and atelectasis at the base of the right lung. The cardiomeastinal silhouette and hilar contours are grossly unchanged. There is no evidence of pneumothorax. Thoracolumbar fusion hardware is seen unchanged in appearance.

Single-image



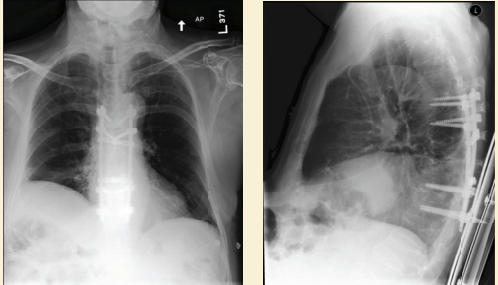
radiology image: <image> Which of the following findings are present in the radiology image? Findings: Enlarged Cardiomeastinum, ...

Support devices.

M4CXR Based on the previous conversation, provide a description of the findings in the radiology image.

M4CXR Frontal and lateral views of the chest were obtained. Left-sided Port-A-Cath is again seen terminating in the low SVC. There is persistent elevation of the right hemidiaphragm. No focal consolidation pleural effusion or evidence of pneumothorax is seen. The cardiac and mediastinal silhouettes are stable. Spinal hardware is again seen.

Multi-image



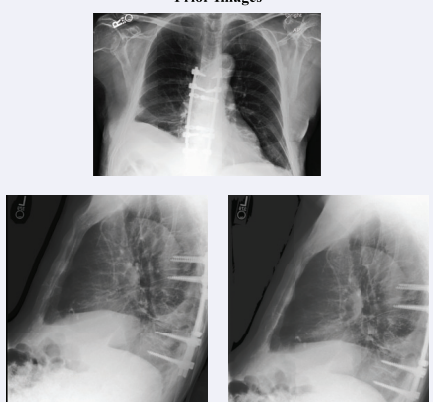
radiology images: <image> <image> Which of the following findings are present in the radiology image? Findings: Enlarged Cardiomeastinum, ...

Atelectasis, Support devices.

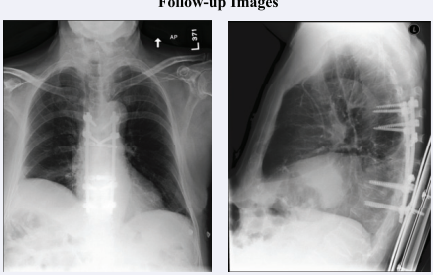
M4CXR Based on the previous conversation, provide a description of the findings in the radiology image.

M4CXR Left-sided Port-A-Cath tip terminates in the mid SVC. The cardiac mediastinal and hilar contours are unchanged. There is no pleural effusion or pneumothorax. Elevation of the right hemidiaphragm is unchanged. There is minimal atelectasis in the right lung base. Spinal fusion hardware is again noted.

Prior Images



Follow-up Images



Multi-study

prior radiology images: <image> <image> <image>, **prior radiology report:** AP and lateral views of the chest. Posterior fixation hardware in the thoracic spine is several years old. Elevation of the right lung base has been increasing slowly over the past years. Small right pleural effusion is comparable to slightly larger than on . Left subclavian line ends in the distal SVC. Heart size is normal. Cardiomeastinal and hilar contours are normal. Right basilar linear atelectasis is unchanged. No consolidation or pulmonary edema is present. **follow-up images:** <image> <image>, The radiology studies are given in chronological order. Which of the following findings are present in the radiology image? Findings: Enlarged Cardiomeastinum, ...

Atelectasis, Pleural effusion, Support devices.

M4CXR Based on the previous conversation, provide a description of the findings in the current follow-up radiology images.

M4CXR Left-sided Port-A-Cath tip terminates in the mid SVC. Thoracic spinal hardware is again noted. The cardiac mediastinal and hilar contours are unchanged. There is no pulmonary vascular congestion. Small right pleural effusion is unchanged. Patchy atelectasis is seen in the right lung base. No pneumothorax is identified. There are no acute osseous abnormalities

Fig. 6. Examples of MRG across various scenarios. For the same study, the top left shows the result for *single-image*, the top right for *multiimage*, and the bottom for *multistudy* report generation. The results are selected from the MIMIC-CXR test set.

TABLE IX

CONVERSATION TEMPLATE FOR EACH TASK. PHRASES INSIDE BRACKETS {} ARE REPLACED WITH APPROPRIATE TEXT BASED ON DATA INSTANCES, WITH RED-COLORED TEXT INDICATING TARGET OUTPUTS. < **IMAGE** > INDICATES THE POSITIONS WHERE VISUAL TOKENS ARE INSERTED

Task	Conversation Template
<i>Single-image</i>	User: radiology image: <image> Which of the following findings are present in the radiology images? Findings: {findings} Assistant: {findings} User: Based on the previous conversation provide a description of the findings in the radiology image. Assistant: {report}
<i>Multi-image</i>	User: radiology images: {images} Which of the following findings are present in the radiology images? Findings: {findings} Assistant: {findings} User: Based on the previous conversation provide a description of the findings in the current follow-up radiology images. Assistant: {report}
<i>Multi-study</i>	User: prior radiology images: {prior images} prior radiology report: {prior report} follow-up images: {follow-up images} The radiology studies are given in chronological order. Which of the following findings are present in the current follow-up radiology images? Findings: {findings} Assistant: {findings} User: Based on the previous conversation provide a description of the findings in the current follow-up radiology images. Assistant: {report}
Disease Classification	User: radiology image: <image> Which of the following findings are present in the radiology image? Findings: {findings} Assistant: {findings}
Finding Grounding	User: radiology image: <image> Is {finding} present in the radiology image? If so, provide the bounding box coordinates of the region. Assistant: {bbox}
Grounded Finding	User: radiology image: <image> Provide a finding name for this region. {bbox} Assistant: {finding}
Abnormality Detection	User: radiology image: <image> Provide the bounding box coordinates of abnormal regions in the radiology image. Assistant: {bbox}
Multi Finding Grounding	User: radiology image: <image> Which of the following findings are present in the radiology image? Provide the bounding box coordinates if present. Findings: {findings} Assistant: {findings, bboxes}
Organ Grounding	User: radiology image: <image> Provide the bounding box coordinates of {organ} in the radiology image. Assistant: {bbox}
Grounded Organ	User: radiology image: <image> Provide an organ name for this region. {bbox} Assistant: {organ}
Grounded Phrase Generation	User: radiology image: <image> Provide a radiology report phrase for the region. {bbox} Assistant: {phrase}
Phrase Grounding	User: radiology image: <image> Provide the bounding box coordinate of the region this phrase describes: {phrase} Assistant: {bbox}
Anatomical Region Grounding	User: radiology image: <image> Provide the bounding box coordinate of the anatomical region. {name} Assistant: {bbox}
Grounded Anatomical Region	User: radiology image: <image> Provide an anatomical region name for this region. {bbox} Assistant: {name}
Visual Question Answering	User: radiology image: <image> Answer the question. {question} Assistant: {answer}
Difference Visual Question Answering	User: reference: <image> main: <image> Using the provided reference and main radiology images answer the following question. {question} Assistant: {answer}
Visual Instruction-following	User: radiology image: <image> {question} Assistant: {answer}

A notable flaw is that expressions such as “increasing” and “worsening,” which would be meaningful only in a *multistudy* scenario, were generated despite the *single-image* context. While it is understandable that this hallucination is due to the ground-truth reports containing such expressions during training, it is evident that there is room for improvement.

To emphasize the impact of incorporating VQA datasets, Fig. 5 illustrates the differences in VQA performance between M4CXR and the Exp4 model in Section V-E, using images selected from the MIMIC-CXR test set. When recommen-

dations for diagnosis and treatment are requested, M4CXR suggests continued observation and follow-up imaging studies based on the findings, and also recommends treatments such as pleural drainage or placement of the chest tube if the condition worsens. In contrast, the Exp4 model provides a simpler response, suggesting a CT evaluation and the need for monitoring. This comparison between the two models demonstrates that VQA datasets play an important role in maintaining the LLM’s ability to provide specific and helpful responses.

TABLE X
NUMBER OF TRAINING INSTANCES AND SAMPLING RATIOS FOR EACH DATASET USED ACROSS TASKS

Task type	Task	Dataset	Train	Dataset ratio	Task type ratio
Medical Report Generation	<i>Single-image</i>	MIMIC-CXR	270,236	200.0	0.54
	<i>Multi-image</i>	MIMIC-CXR	151,606	120.0	
	<i>Multi-study</i>	MIMIC-CXR	68,373	60.0	
Image Understanding	Disease Classification	BRAX	40,965	7.0	0.35
		CheXpert	223,414	30.0	
		VinDr-CXR	15,000	3.0	
		ChestX-ray14	86,523	10.0	
		ChestX-Det10	3,578	1.0	
		SIIM	10,675	4.0	
		RSNA	14,863	4.0	
		COVID-19 Radiography	15,153	3.0	
	Finding Grounding	VinDr-CXR	15,000	2.0	
		ChestX-ray14	50,500	10.0	
		ChestX-Det10	3,578	2.0	
		JSRT	247	0.1	
		SIIM	10,675	4.0	
		RSNA	14,863	1.0	
		COVID-QU-Ex	2,796	1.0	
		QaTa-COV19	4,194	1.0	
	Grounded Finding	VinDr-CXR	4,394	2.0	
		ChestX-Det10	2,967	2.0	
		JSRT	154	0.1	
		SIIM	2,379	2.0	
RSNA		6,012	1.0		
COVID-QU-Ex		1,864	1.0		
QaTa-COV19		4,194	3.0		
Abnormality Detection		VinDr-CXR	4,394	3.0	
	ChestX-Det10	2,967	1.5		
	JSRT	154	0.1		
	SIIM	2,379	2.0		
	RSNA	6,012	1.0		
	COVID-QU-Ex	1,864	1.5		
	QaTa-COV19	4,194	1.5		
	Multi Finding Grounding	ChestX-ray14	50,500	10.0	
VinDr-CXR		15,000	2.0		
ChestX-Det10		3,578	2.0		
Organ Grounding	COVID-19 Radiography	15,153	0.8		
	COVID-QU-Ex	3,728	0.8		
Grounded Organ	COVID-19 Radiography	15,153	0.8		
	COVID-QU-Ex	3,728	0.8		
Grounded Phrase Generation	MS-CXR	638	2.0		
	ImaGenome	164,229	40.0		
Phrase Grounding	MS-CXR	638	2.0		
	ImaGenome	164,229	40.0		
Anatomical Region Grounding	ImaGenome	164,229	20.0		
	Grounded Anatomical Region	ImaGenome	164,229	20.0	
Visual Question Answering	Visual Question Answering	MIMIC-CXR-VQA	255,919	2.0	0.11
		MIMIC-Diff-VQA	553,156	2.0	
	Difference Visual Question Answering	MIMIC-Diff-VQA	129,900	4.0	
	Visual Instruction-following	RaDialog	297,964	70.0	

In Fig. 5, the medical report generated by M4CXR contains comparative contents such as “In comparison with study” and “decreased,” even though it was generated in a *single-image* scenario. Similar issues can be observed in other examples, where the model often refers to comparisons or mentions images that are not provided. During the training of *single-image* MRG, the use of certain ground-truth reports, those written specifically in a *multistudy* context, is likely to have induced these hallucinations. To address this issue, ground-truth reports tailored to each MRG

scenario are needed, which could be explored in future work.

Fig. 6 presents the results of report generation for the same study across various scenarios, highlighting how clinical accuracy improves as more information becomes available. The ground-truth targets, including atelectasis, pleural effusion, and support devices, are each marked in different colors. In *single image*, one lateral image is used; in *multiimage*, both AP and lateral images are used; and in *multistudy*, three prior images and the corresponding report are included.

In *single image*, only the support devices are correctly identified, suggesting that a single lateral image may not provide sufficient information for accurate CXR interpretation. In *multiimage*, the addition of a frontal image enables the correct identification of atelectasis. In *multistudy*, all three targets are successfully identified, possibly due to the prior report already containing these three findings. This demonstrates that M4CXR is capable of understanding and utilizing the information provided in various MRG scenarios.

VI. CONCLUSION

In this study, we introduced M4CXR, a multimodal LLM aimed at enhancing CXR interpretation by leveraging the versatile advantages of LLMs. Trained on a CXR visual instruction-following dataset constructed by appropriately combining various datasets, M4CXR is capable of performing multiple tasks. Exploiting its reasoning capability, we proposed a novel CoT prompting strategy that significantly improved the clinical accuracy of report generation. By identifying observations from X-ray images and generating descriptions through multiturn conversations, M4CXR achieved notable improvements in CheXbert *F1*-scores compared to existing models. The model’s ability to handle multiple images and incorporate prior studies demonstrated its adaptability in diverse clinical scenarios. Beyond report generation, M4CXR also proved to be highly effective in visual grounding and VQA. M4CXR competed well with models specialized for visual grounding, while also exhibiting outstanding performance in VQA. Qualitative analysis highlights M4CXR’s flexibility in answering free-form questions.

Although M4CXR demonstrated strong performance in CheXbert metrics, it also exhibited clear limitations, such as low RadGraph *F1*-scores and the occurrence of hallucinations. Some potential improvements include incorporating the concept of knowledge graphs to improve RadGraph results or using LLMs for data preprocessing to eliminate hallucinations. With future work to address these limitations, we anticipate that further advancements in M4CXR will contribute to the development of a highly capable radiology assistant.

APPENDIX A DATASET PREPROCESSING

We follow the official split for every dataset. The additional preprocessing steps, taken for each dataset as required, are described in detail. For MIMIC-CXR, we extract the FINDINGS section using the official preprocessing code.¹ Then, we use the text preprocessing code from [62] with slight modification to remove special characters, numbering, and extra spaces. We exclude reports that lack a FINDINGS section or have a length of less than five characters. For *multiimage*, we consider only studies with at most five images. For *multistudy*, we include only cases where the combined number of images in the prior and follow-up studies is at most 10. Out of 2461 follow-up studies in the test set, only 2333 studies have a prior study, and the remaining 128 studies have no preceding study.

Especially, image understanding datasets that involve mask labels require extra procedures. Due to the presence of many similar bounding box labels in VinDR-CXR, overlapping boxes in an image are merged into the smallest possible bounding box if they overlap by more than 50%. Since the labels of JSRT are circular, they are converted to the smallest possible bounding box labels. The mask labels of ChestX-Det10, SIIM, COVID-19 Radiography, COVID-QU-EX, and QaTa-COV19 are converted to bounding boxes that minimally enclose each mask. For COVID-19 Radiography, images with more than three mask regions are excluded, and data in QaTa-COV19 that overlaps with COVID-QU-EX are removed. We utilize the code from [17] to preprocess ImaGenome, obtaining report text, bounding boxes, and anatomical names for 29 anatomical regions. Additionally, among the disease labels of RSNA dataset, only “lung opacity” and “normal” are used, and according to [63], “lung opacity” is referred to as “pneumonia.”

For RaDialog, we use only the images from MIMIC-CXR. Among the tasks introduced in [6], we exclude RG (report generation). We utilize version 1.1.0 [60] of the dataset.

APPENDIX B DETAILS ON CXR INTERPRETATION TASKS

Table IX provides the conversation templates for each task. Although not shown in the table, a system prompt, which specifies that the AI medical assistant should give helpful and detailed answers, is inserted before the first question in each template. Table X lists the datasets used for each task, along with the number of training instances and the sampling ratios applied during training.

REFERENCES

- [1] Y. Zheng, W. Gan, Z. Chen, Z. Qi, Q. Liang, and P. S. Yu, “Large language models for medicine: A survey,” 2024, *arXiv:2405.13055*.
- [2] Z. Zhou, M. Shi, M. Wei, O. Alabi, Z. Yue, and T. Vercauteren, “Large model driven radiology report generation with clinical quality reinforcement learning,” 2024, *arXiv:2403.06728*.
- [3] S. L. Hyland et al., “MAIRA-1: A specialised large multimodal model for radiology report generation,” 2023, *arXiv:2311.13668*.
- [4] J. M. Z. Chaves et al., “Towards a clinically accessible radiology foundation model: Open-access and lightweight, with automated evaluation,” 2024, *arXiv:2403.08002*.
- [5] O. Thawakar et al., “XrayGPT: Chest radiographs summarization using medical vision-language models,” 2023, *arXiv:2306.07971*.
- [6] C. Pellegrini, E. Özsoy, B. Busam, N. Navab, and M. Keicher, “RaDialog: A large vision-language model for radiology report generation and conversational assistance,” 2023, *arXiv:2311.18681*.
- [7] S. Lee, W. Jun Kim, J. Chang, and J. Chul Ye, “LLM-CXR: Instruction-finetuned LLM for CXR image understanding and generation,” 2023, *arXiv:2305.11490*.
- [8] Z. Chen et al., “A vision-language foundation model to enhance efficiency of chest X-ray interpretation,” 2024, *arXiv:2401.12208*.
- [9] J. Wei et al., “Chain-of-thought prompting elicits reasoning in large language models,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24824–24837.
- [10] C. Raffel et al., “Exploring the limits of transfer learning with a unified text-to-text transformer,” *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.
- [11] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, Jan. 2023, pp. 34892–34916.
- [12] J.-B. Alayrac et al., “Flamingo: A visual language model for few-shot learning,” in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, Jan. 2022, pp. 23716–23736.

¹<https://github.com/MIT-LCP/mimic-cxr/tree/master/txt>

- [13] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," 2023, *arXiv:2304.10592*.
- [14] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, Jan. 2023, pp. 19730–19742.
- [15] J. Cha, W. Kang, J. Mun, and B. Roh, "Honeybee: Locality-enhanced projector for multimodal LLM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 13817–13827.
- [16] H.-C. Shin, K. Roberts, L. Lu, D. Demner-Fushman, J. Yao, and R. M. Summers, "Learning to read chest X-rays: Recurrent neural cascade model for automated image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2497–2506.
- [17] T. Tanida, P. Müller, G. Kaissis, and D. Rueckert, "Interactive and explainable region-guided radiology report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7433–7442.
- [18] M. Yamaç, M. Ahishali, A. Degerli, S. Kiranyaz, M. E. H. Chowdhury, and M. Gabbouj, "Convolutional sparse support estimator-based COVID-19 recognition from X-ray images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 1810–1820, May 2021.
- [19] Z. Lin et al., "AANet: Adaptive attention network for COVID-19 detection from chest X-ray images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 11, pp. 4781–4792, Nov. 2021.
- [20] H. Jin, H. Che, Y. Lin, and H. Chen, "PromptMRG: Diagnosis-driven prompts for medical report generation," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2024, vol. 38, no. 3, pp. 2607–2615. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/28038>
- [21] H. Li, H. Wang, X. Sun, H. He, and J. Feng, "Prompt-guided generation of structured chest X-ray report using a pre-trained LLM," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2024, pp. 1–6.
- [22] H. Li, H. Wang, X. Sun, H. He, and J. Feng, "Context-enhanced framework for medical image report generation using multimodal contexts," *Knowl.-Based Syst.*, vol. 310, Feb. 2025, Art. no. 112913.
- [23] L. Yang et al., "Advancing multimodal medical capabilities of Gemini," 2024, *arXiv:2405.03162*.
- [24] T. Tu et al., "Towards generalist biomedical AI," *NEJM AI*, vol. 1, no. 3, Feb. 2024, Art. no. 2300138.
- [25] B. Boecking et al., "Making the most of text semantics to improve biomedical vision-language processing," in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, Jan. 2022, pp. 1–21.
- [26] Z. Chen et al., "Medical phrase grounding with region-phrase context contrastive alignment," in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent. Cham, Switzerland: Springer*, Jan. 2023, pp. 371–381.
- [27] S. Bannur et al., "MAIRA-2: Grounded radiology report generation," 2024, *arXiv:2406.04449*.
- [28] H. You et al., "Ferret: Refer and ground anything anywhere at any granularity," in *Proc. 12th Int. Conf. Learn. Represent.*, Jan. 2023.
- [29] J. Chen et al., "MiniGPT-v2: Large language model as a unified interface for vision-language multi-task learning," 2023, *arXiv:2310.09478*.
- [30] J. Irvin et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," 2019, *arXiv:1901.07031*.
- [31] A. E. W. Johnson et al., "MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs," 2019, *arXiv:1901.07042*.
- [32] A. Smit, S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. P. Lungren, "CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT," 2020, *arXiv:2004.09167*.
- [33] E. P. Reis et al., "BRAX, Brazilian labeled chest X-ray dataset," *Scientific Data*, vol. 9, no. 1, p. 487, Aug. 2022.
- [34] H. Q. Nguyen et al., "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," 2020, *arXiv:2012.15029*.
- [35] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 3462–3471.
- [36] J. Liu, J. Lian, and Y. Yu, "ChestX-det10: Chest X-ray dataset on detection of thoracic abnormalities," 2020, *arXiv:2006.10550*.
- [37] J. Shiraishi et al., "Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of Radiologists' detection of pulmonary nodules," *Amer. J. Roentgenology*, vol. 174, no. 1, pp. 71–74, Jan. 2000.
- [38] A. Zawacki et al. (2019). *SIIM-ACR Pneumothorax Segmentation*. [Online]. Available: <https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation>
- [39] G. Shih et al., "Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia," *Radiol., Artif. Intell.*, vol. 1, no. 1, Jan. 2019, Art. no. e180041.
- [40] M. E. H. Chowdhury et al., "Can AI help in screening viral and COVID-19 pneumonia?," *IEEE Access*, vol. 8, pp. 132665–132676, 2020.
- [41] T. Rahman et al., "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-rays images," *Comput. Biol. Med.*, vol. 2020, Jan. 2020, Art. no. 104319.
- [42] A. Tahir et al. (2021). *COVID-Qu-EX Dataset*. [Online]. Available: <https://www.kaggle.com/datasets/tahirahmed/covidquex>
- [43] A. Degerli et al., "COVID-19 infection map generation and detection from chest X-ray images," *Health Inf. Sci. Syst.*, vol. 9, no. 1, p. 15, Dec. 2021.
- [44] J. T. Wu et al., "Chest ImaGenome dataset for clinical reasoning," 2021, *arXiv:2108.00316*.
- [45] S. Bae et al., "EHRXQA: A multi-modal question answering dataset for electronic health records with chest X-ray images," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 36, Jan. 2023, pp. 3867–3880.
- [46] X. Hu et al., "Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2023, pp. 4156–4165, doi: [10.1145/3580305.3599819](https://doi.org/10.1145/3580305.3599819).
- [47] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, "SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering," 2021, *arXiv:2102.09542*.
- [48] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, Philadelphia, PA, USA: Association for Computational Linguistics, 2002, pp. 311–318.
- [49] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, Apr. 2004, pp. 74–81.
- [50] A. Nicolson, J. Dowling, and B. Koopman, "Improving chest X-ray report generation by leveraging warm starting," *Artif. Intell. Med.*, vol. 144, Oct. 2023, Art. no. 102633.
- [51] F. Yu et al., "Evaluating progress in automatic chest X-ray radiology report generation," *Patterns*, vol. 4, no. 9, Sep. 2023, Art. no. 100802.
- [52] S. Jain et al., "RadGraph: Extracting clinical entities and relations from radiology reports," in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, J. Vanschoren and S. Yeung, Eds., Jan. 2021. [Online]. Available: https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/c8ffe9a587b126f152ed3d89a146b445-Paper-round1.pdf
- [53] A. Q. Jiang et al., "Mistral 7B," 2023, *arXiv:2310.06825*.
- [54] F. Pérez-García et al., "Exploring scalable medical image encoders beyond text supervision," 2024, *arXiv:2401.10815*.
- [55] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.
- [56] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," 2023, *arXiv:2307.08691*.
- [57] Z. Wang, L. Liu, L. Wang, and L. Zhou, "METransformer: Radiology report generation by transformer with multiple learnable expert tokens," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 11558–11567.
- [58] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest X-ray report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3334–3343.
- [59] J. Deng, Z. Yang, T. Chen, W. Zhou, and H. Li, "TransVG: End-to-end visual grounding with transformers," 2021, *arXiv:2104.08541*.
- [60] C. Pellegrini, E. Özsoy, B. Busam, N. Navab, and M. Keicher, "Radialog instruct dataset (version 1.1.0)," *PhysioNet*, 2024, doi: [10.13026/40d2-wj86](https://doi.org/10.13026/40d2-wj86).
- [61] C. Wu, X. Zhang, Y. Zhang, Y. Wang, and W. Xie, "Towards generalist foundation model for radiology by leveraging web-scale 2D & 3D medical data," 2023, *arXiv:2308.02463*.
- [62] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," 2020, *arXiv:2010.16056*.
- [63] K. You et al., "CXR-clip: Toward large scale chest X-ray language-image pre-training," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2023*. Cham, Switzerland: Springer, 2023, pp. 101–111.



Jongwon Park received the B.S. and Ph.D. degrees in industrial engineering from Seoul National University (SNU), Seoul, South Korea, in 2018 and 2023, respectively.

He is currently an AI Researcher at DEEPNOID Inc., Seoul. His research interests include vision-language models, multimodal large language models, multimodal representation learning, and medical imaging.



Jihun Hyun received the B.S. degree in electrical and electronic engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2017, and the M.S. degree in computer science and engineering from Ulsan National Institute of Science and Technology (UNIST), Ulsan, South Korea, in 2019.

He is currently the Head of the AI Research Laboratory at DEEPNOID Inc., Seoul, South Korea. His research interests include artificial intelligence, multimodal vision-language models, and heterogeneous systems.



Soobum Kim received the B.S. and M.S. degrees in biomedical engineering from Gachon University, Seongnam, South Korea, in 2018 and 2020, respectively.

He worked as a Research Engineer at the AI Research Laboratory, DEEPNOID Inc., Seoul, South Korea. His research interests include artificial intelligence, medical data analysis, and circuit design.



Byungmu Yoon received the B.S. degree in biomedical engineering from Gachon University, Seongnam, South Korea, in 2020.

He is currently working as a Research Engineer at the AI Research Laboratory, DEEPNOID Inc., Seoul, South Korea. His research interests include artificial intelligence, medical imaging, and multimodal vision-language models.



Kyoyun Choi received the B.S. and Ph.D. degrees in industrial engineering from Seoul National University (SNU), Seoul, South Korea, in 2016 and 2021, respectively.

He worked as a Post-Doctoral Researcher at the Institute of Engineering Research, SNU, from 2021 to 2022. He is currently working as a Research Engineer at the AI Research Laboratory, DEEPNOID Inc., Seoul. His research interests include artificial intelligence, generative modeling, medical imaging, and multimodal vision-language models.