WoLF: Wide-scope Large Language Model Framework for CXR Understanding

Seil Kang[1], Donghyun Kim[1], Junhyeok Kim[1], Hyo Kyung Lee[2], Seong Jae Hwang[1]

[1]Yonsei University, [2]Korea University
{seil, danny0103, timespt, seongjae}@yonsei.ac.kr,
hyokyunglee@korea.ac.kr

Abstract. Significant methodological strides have been made toward Chest X-ray (CXR) understanding via modern vision-language models (VLMs), demonstrating impressive Visual Question Answering (VQA) and CXR report generation abilities. However, existing CXR understanding frameworks still possess several procedural caveats. (1) Previous methods solely use CXR reports, which are insufficient for comprehensive Visual Question Answering (VQA), especially when additional health-related data like medication history and prior diagnoses are needed. (2) Previous methods use raw CXR reports, which are often arbitrarily structured. While modern language models can understand various text formats, restructuring reports for clearer, organized anatomy-based information could enhance their usefulness. (3) Current evaluation methods for CXR-VQA primarily emphasize linguistic correctness, lacking the capability to offer nuanced assessments of the generated answers. In this work, to address the aforementioned caveats, we introduce **WoLF**, a **W**ide-sc**o**pe **L**arge Language Model **F**ramework for CXR understanding. To resolve (1), we capture multi-faceted records of patients, which are utilized for accurate diagnoses in real-world clinical scenarios. Specifically, we adopt the Electronic Health Records (EHR) to generate instruction-following data suited for CXR understanding. Regarding (2), we enhance report generation performance by decoupling knowledge in CXR reports based on anatomical structure even within the attention step via masked attention. To address (3), we introduce an AI-evaluation protocol optimized for assessing the capabilities of LLM. Through extensive experimental validations, WoLF demonstrates superior performance over other models on MIMIC-CXR in the AI-evaluation arena about VQA (up to +9.47%p mean score) and by metrics about report generation (+7.3%p BLEU-1).

Keywords: CXR Understanding · LLM Framework · Instruction Tuning

1 Introduction

Recent years have witnessed significant progress in the field of Chest X-ray (CXR) understanding, particularly through downstream tasks like Visual Question Answering (VQA) and automated report generation. Despite considerable advancements, we raise issues that models engaged in Chest X-ray (CXR) understanding persistently encounter several challenges from a framework standpoint. (1) Existing approaches [13,24] predominantly depend on CXR reports

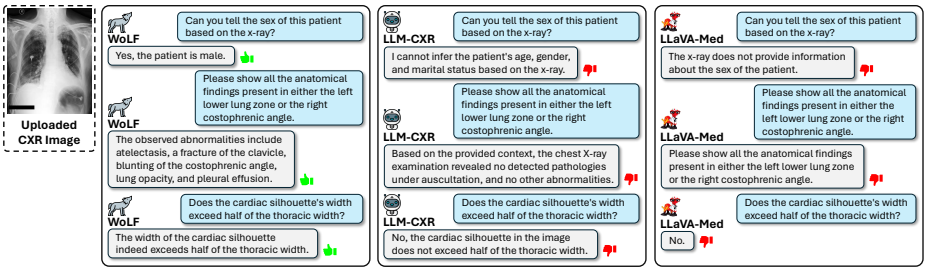


Fig. 1. Comparisons with other models for VQA scenario given a CXR image. Green thumbs indicate the quality of the response is good (accurate, helpful), while red thumbs indicate bad (inaccurate, evasive), with respect to target answers.

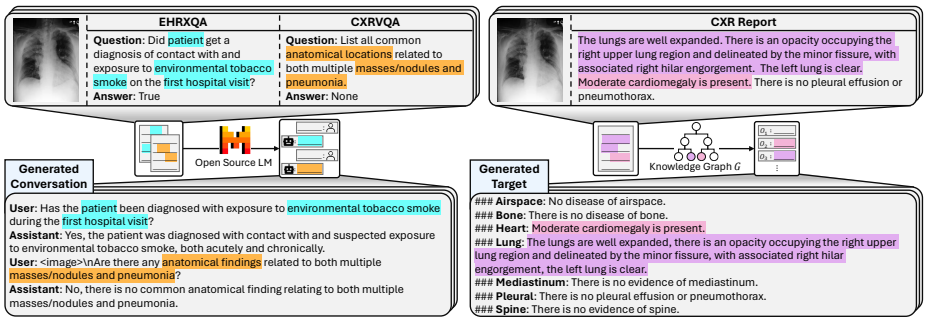
for supervised learning, overlooking the crucial aspect of incorporating patients' personalized health records, which are diagnoses-supportive in real-world clinical scenarios. (2) Additionally, the performance of report generation is constrained by the unstructured format of CXR reports. Unstructured raw CXR reports, exemplified by Fig. 2(b), impede the ability of models to learn CXR anatomical structures in supervised learning settings, owing to their non-intuitive format. (3) Lastly, the existing evaluation metrics for CXR-VQA primarily focus on the correctness of answers, which falls short in assessing the generative language models' comprehensive understanding of CXR imagery.

To tackle the issues illustrated above, we introduce WoLF, a **W**ide-sc**o**pe **L**arge Language Model **F**ramework for CXR understanding. We will delve into the specifics of our approach, detailing the innovative solutions we develop for each challenge:

(1) For more in-depth use of such systems in practice, as exemplified in Fig. 1, the model must consider various patient records, including Electronic Health Records (EHR). Thus, we hypothesize that incorporating patients' personalized EHR records can enhance the CXR understanding of vision-language models. To validate this hypothesis, we introduce *Health-specific Instruction Tuning (HIT)* to deal with the existing limitations that training merely relies on CXR reports.

(2) Unorganized CXR reports restrict the advancement in report generation tasks. To push the envelope, we present *Anatomy-Specific Knowledge decoupling (ASK)* to separate the reports into anatomy-specific findings. The generated targets give a model a direct understanding of a specific anatomical structure, without being disturbed by other structures. Synchronized with ASK, we introduce *Anatomy-localizing Masked Attention (AMA)* that promotes independent learning on each anatomical structure.

(3) Current evaluation methods for CXR-VQA mostly emphasize linguistic correctness. These methods are incapable of assessing the responses from generative language models across a wide range. Inspired by [12,19], we provide a novel *AI-evaluation protocol* that is well-suited to generative language models across dimensions of *Accuracy, Helpfulness, Relevance, Hallucination*, and *Universality*. Through our extensive AI evaluation, we can discern the extent to which



(a) Health-specific Instruction Tuning **(b) Anatomy-Specific Knowledge decoupling**

Fig. 2. Data generation overview of HIT and ASK: (a) We generate health-specific instruction-following dataset. In (a), Cyan and orange sequences are queries about EHR and findings in CXR respectively. (b) We reorganize original CXR reports into sequences of anatomy-specific structures through the use of a knowledge graph, G.

models understand CXR from their VQA results, rather than just evaluating the correctness of the models' responses.

To sum up, the contribution of our model can be described at the *macro* and *micro* level, respectively; Macroscopically, our framework covers data reformulation, training method to improve CXR understanding, and AI-evaluation protocol. Microscopically, **(i)** we present a novel instruction-following data tuning method called Health-specific Instruction Tuning (HIT) designed for interplay between personalized health records and visual representations of CXR. **(ii)** We propose Anatomy-Specific Knowledge decoupling (ASK), for hierarchically breaking down a radiology report by anatomical structures. Furthermore, we present Anatomical-localizing Masked Attention to support the merits of decoupled data from ASK, enabling expertised visual-language comprehension for each anatomical structure. **(iii)** As the final step of the framework, we introduce AI-evaluation for advanced analysis of our model. This evaluates the broad capabilities of generative language models on the VQA task. **(iv)** Through these methods, our study achieved state-of-the-art performance in the report generation and VQA tasks on MIMIC-CXR [10] and IU-Xray [7].

2 WoLF: Wide-scope Large Language Model Framework

We introduce **W**ide-sc**o**pe **L**arge Language Model **F**ramework (WoLF). WoLF establishes its framework through the following macro-level steps:

$$\underset{\text{(Sec. 2.1)}}{\text{Data Reformulation}} \rightarrow \underset{\text{(Sec. 2.2)}}{\text{Model Training}} \rightarrow \underset{\text{(Sec. 2.3)}}{\text{AI-evaluation}}$$

In Sec. 2.1, we describe our innovative data reformulation scheme designed for VQA and report generation, the two main CXR understanding tasks. In Sec. 2.2,

Algorithm 1 HIT: Health-specific Instruction Tuning

Input: EHR QA-set: $\{q^{\text{ehr}}, a^{\text{ehr}}\}$, VQA QA-set: $\{q^{\text{vqa}}, a^{\text{vqa}}\}$, # of {Patients, Studies}: $\{P, N_p\}$, Open-source Language Model: f_ϕ

global $S \leftarrow$ `SystemPrompt`
for $p = 1$ to P **do**
  for $i = 1$ to N_p **do**
    $Q \leftarrow (q_{pi}^{\text{ehr}}, q_{pi}^{\text{vqa}})$
    $A \leftarrow (a_{pi}^{\text{ehr}}, a_{pi}^{\text{vqa}})$
    $S_i \leftarrow$ PromptGenerator(Q, A, f_ϕ)
    $S \leftarrow$ concatenate(S, S_i)
Output: Generated S

function PromptGenerator(Q, A, f)
  for q_t, a_t in $\{Q, A\}$ **do**
    $S \leftarrow$ concatenate(S, q_t, a_t)
  return $f(S)$

Algorithm 2 ASK: Anatomy-specific Knowledge decoupling

Input: Knowledge graph: G, Anatomical structures: O, # of {Patients, Studies, Anatomical structures}: $\{P, N_p, M\}$, CXR Reports: R

$D^* \leftarrow \emptyset$
Def. 1: $R = \{r_i | i = 1, 2, \ldots, N_p\}$
for $p = 1$ to P **do**
  for $i = 1$ to N_p **do**
    Append Decoupler(O, G, r_i) to D^*
Output: Decoupled D^*

function Decoupler(O, G, r)
  Def. 2: $O = \{o_m | m = 1, 2, \ldots, M\}$
  $D \leftarrow$ `EmptyString("")`
  $T \leftarrow$ `AnatomyTag`
  for $m = 1$ to M **do**
    $D \leftarrow$ concatenate$(D, T, G(r, o_m))$
  return D

we introduce a *two-stage* training approach to further enhance the model's performance in these tasks. Finally, in Sec. 2.3, we propose an AI-evaluation protocol tailored for assessing generative language models in the CXR-VQA.

2.1 Data Reformulation

To train WoLF as a framework that excels in performing VQA and report generation tasks for CXR, we propose two data reformulation methods: ***H**ealth-specific **I**nstruction **T**uning (**HIT**)* for VQA, and ***A**natomy-**S**pecific **K**nowledge decoupling (**ASK**)* for report generation.

HIT. LLaVA [18] and LLaVA-Med [14] have shown the effectiveness of explicitly tuned data combined with vision-language *instruction tuning*, enabling thorough reasoning about vision-language data. Specifically, to encode images into their visual features, GPT-4 [21] is used as a text-only input teacher. The newly generated data from the teacher takes the form of a multi-turn conversation between the `USER` and the `ASSISTANT`. Inspired by LLaVA [18], our approach, HIT (Alg. 1), focuses on CXR understanding, in particular, CXR VQA. That is, we generate a novel health-specific instruction-following dataset of multi-turn conversations that comprise EHR and the CXR findings. For instance,

`User`: Has patient been diagnosed with contact or exposure to tobacco **smoke**?
`Assistant`: Yes, the patient has been exposured with **smoking environment**.
`User`: <IMG> Is there any sign of the **pneumonia** in the right apical zone?
`Assistant`: Yes, the image has evidence of **pneumonia** in right apical zone.

<IMG> is a placeholder for image embeddings. HIT employs dataset [2] constructed from MIMIC-IV [9] and MIMIC-CXR [10].

ASK. During training, models struggle to recognize individual anatomical structures in unstructured CXR reports. For better report generation, we need CXR

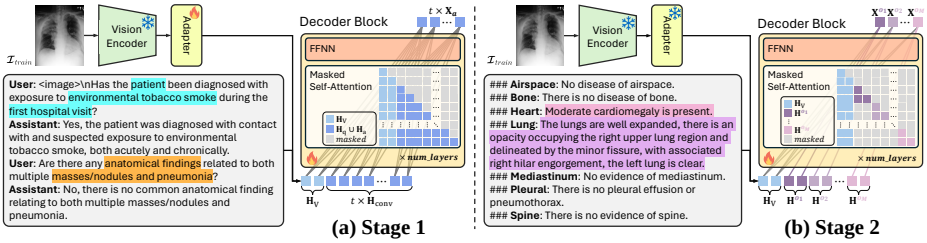


Fig. 3. Overview of the training phase. (a) The input embedding $\mathbf{H}_{(\cdot)}$ consists of visual embedding from the adapter and language embeddings from t-turn conversations generated by HIT. Cyan and orange sequences are queries about EHR and findings in CXR respectively. (b) For an anatomical structure o_m, its sequence embedding is denoted by $\mathbf{H}^{o_m}$. Organized CXR reports by ASK are utilized as input for training.

reports organized by anatomy in supervised learning. To this end, we present ASK to reorganize the CXR reports for more accurate report generation. As explicated in Zhang et al. [30], attributes in the CXR reports are classified based on the knowledge graph G. We separate sentences of CXR reports based on anatomy-specific diseases (e.g., `{pneumonia, edema}`→`lung`). Eventually, we obtain target data consisting of `AnatomyTag` and its findings (e.g., `### Heart: Moderate cardiomegaly is present.`). In the prior study, ITA [25] trains decoder heads to generate sentences for anatomical structures and aggregates training losses. In contrast, we directly dissociate the CXR report into independent anatomical sentences (Alg. 2). Next, we describe how we use this refined CXR instruction data to train WoLF.

2.2 Model Training

The training phase of WoLF is divided into two stages (Fig. 3). In stage 1, the model is trained specifically for the VQA task, while stage 2 focuses on the report generation task, leveraging the CXR understanding trained in the previous stage. Thus, WoLF holistically allows both VQA and report generation based on its superior CXR understanding.

Stage 1. The proposed dataset comprises the following components at the token level; images $\mathbf{X}_\mathrm{V}$, system prompt $\mathbf{X}_\mathrm{sys}$, user questions $\mathbf{X}_\mathrm{q}$, and assistant answers $\mathbf{X}_\mathrm{a}$. Precisely, for a sequence of length L, we compute the probability of the assistant answers $\mathbf{X}_\mathrm{a}$ as target answers by

$$p(\mathbf{X}_\mathrm{a}|\mathbf{X}_\mathrm{V},\mathbf{X}_\mathrm{sys},\mathbf{X}_\mathrm{q}) = \prod_{i=1}^{L} p_\theta(x_{\mathrm{a},i}|\mathbf{X}_\mathrm{V},\mathbf{X}_\mathrm{sys},\mathbf{X}_{\mathrm{q},<i},\mathbf{X}_{\mathrm{a},<i}), \tag{1}$$

where θ represents trainable parameters. This outlines the objective of the WoLF training phase through HIT instructions. The model seeks to optimize the probability p_θ of predicting the current token $x_{\mathrm{a},i}$, conditioned on $\mathbf{X}_\mathrm{V},\mathbf{X}_\mathrm{sys},\mathbf{X}_\mathrm{q}$, and $\mathbf{X}_{\mathrm{a},<i}$ which denotes the preceding answer tokens before $x_{\mathrm{a},i}$. As shown in

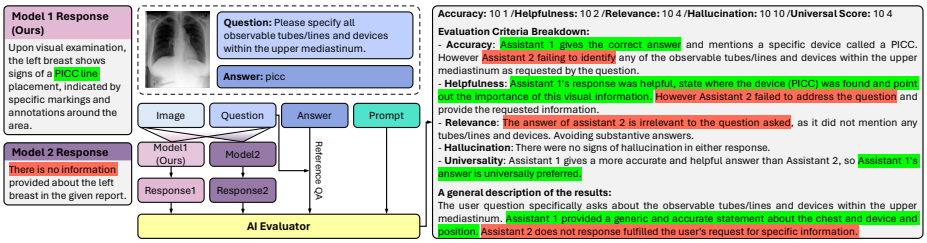


Fig. 4. Our AI-evaluation protocol overview. Green and red are positive and negative feedback from the evaluator, respectively. Only our predictions received positive feedback.

Fig. 2(a), adapter and LLM are trainable but the image encoder is non-trainable. **Stage 2.** To effectively utilize decoupled CXR reports from ASK, we present Anatomy-localizing Masked Attention for the second training phase. We define o_m to be a specific anatomical structure of m^{th}. As shown in Fig. 3(b), when the model predicts $\mathbf{X}^{o_m}$, it attends exclusively on $\mathbf{H}_{\text{q}}^{o_m}$, $\mathbf{X}_{\text{V}}$, and $\mathbf{X}_{\text{sys}}$. The prediction process of our model can be described in an auto-regressive manner, as follows:

$$p(\mathbf{X}^{o_m}|\mathbf{X}_{\text{V}}, \mathbf{X}_{\text{sys}}) = \prod_{i=1}^{L^{o_m}} p_\theta(x_i^{o_m}|\mathbf{X}_{\text{V}}, \mathbf{X}_{\text{sys}}, \mathbf{X}_{<i}^{o_m}), \tag{2}$$

which describes the objective of a model during the second stage for a sequence of o_m. The model predicts L^{o_m} tokens for the sequence of o_m during training. As a result, we focus on guiding the model to learn independent anatomical comprehension in a CXR. Note that we use the final model for all task inferences.

2.3 AI-evaluation

We now present a quantitative analysis of LLMs through AI-evaluation in CXRVQA [2] exploiting prior studies [12,19]. Traditional evaluations, particularly in Visual Question Answering (VQA), have focused merely on correctness. However, we measure five metrics (*Accuracy, Helpfulness, Relevance, Hallucination, Universality*) within the external AI evaluator [1,23]. As shown in Fig. 4, both our model and another comparable model are evaluated by mirroring the human cognitive process, employing a *prompt* detailed in Supp. Fig. 1. Both models generate responses based on the input CXR and visual question. The evaluator receives text-only inputs which are questions, answers, responses from ours and a comparable model, and a *prompt*. The judgment, in Fig. 4, is a human-like evaluation through obvious criteria. To offset position bias, we average the rating for the evaluation set and that for the position-swapped set to get the final rating.

Table 1. LLM capability comparison using relative scores on CXRVQA [2] *test-set*. The scores (*mean*(±*std*)) are converted to a *100-point scale*.

Method	Gemini [23]					
	Accuracy↑	**Helpfulness↑**	**Relevance↑**	**Hallucination↑**	**Universality↑**	**Mean↑**
LLaVA-Med [14]	64.32(±0.35)	66.29(±1.02)	67.32(±1.11)	71.97(±2.43)	67.12(±0.81)	67.41(±0.25)
XrayGPT [24]	30.75(±0.79)	34.59(±0.43)	38.19(±0.78)	55.11(±3.16)	38.85(±0.82)	39.50(±0.41)
LLM-CXR [13]	61.20(±0.97)	69.67(±0.32)	69.29(±1.09)	77.88(±0.73)	67.56(±0.29)	69.12(±0.37)
WoLF	**73.03**(±0.64)	**73.19**(±0.87)	**72.34**(±0.42)	**79.02**(±0.67)	**75.95**(±0.49)	**74.70**(±0.34)
	PaLM-2 [1]					
	Accuracy↑	**Helpfulness↑**	**Relevance↑**	**Hallucination↑**	**Universality↑**	**Mean↑**
LLaVA-Med [14]	67.38(±0.15)	67.35(±0.76)	66.47(±2.30)	72.51(±4.33)	67.51(±0.36)	68.25(±0.31)
XrayGPT [24]	33.54(±0.59)	45.04(±0.30)	36.43(±0.65)	51.11(±0.91)	40.94(±1.10)	41.41(±0.25)
LLM-CXR [13]	68.73(±0.19)	70.03(±1.39)	69.91(±1.95)	74.80(±3.53)	69.80(±0.07)	70.66(±0.06)
WoLF	**79.54**(±0.77)	**79.93**(±1.22)	**79.21**(±0.58)	**85.30**(±3.74)	**78.18**(±0.91)	**80.13**(±1.17)

3 Experiments

We show the VQA results using AI-evaluation and report generation performance of WoLF and other existing methods. Further qualitative results, ablations, and experiments on the same dataset [10] can be found in Supp. Fig. 2.

Training Details. `Vicuna-7b` [6] is used for core LLM with LoRA [8] fine-tuning. CLIP-ViT-L-14 [22] is used as an image encoder. The batch size is set to 64. The learning rate is set to 2e-5. Each stage is trained for 2 epochs.

VQA performance based on AI-evaluation. We use CXRVQA [2] test dataset to show AI-evaluation results. The test dataset comprises 11,309 VQAs for CXR mainly focusing on findings and visual questions that require interpreting images. We analyze with comparable existing models of which the source code is *publicly available and reproducible*. We ensure adherence to domain consistency since the CXRVQA is derived from the MIMIC-CXR [10]. In Table 1, we evaluate scores for each component (*Accuracy, Helpfulness, Relevance, Hallucination, Universality*) to Gemini [23] and PaLM-2 [1] by querying them 4-times each while swapping answer positions (temperature $\tau = 1.0$). Moreover, inspired by RLAIF [12], we instruct the AI evaluator to measure the *win-rate* using a modified prompt referred to Supp. Fig. 1. Specifically, let a visual question set be $\{q_i | i = 1, 2, \ldots, n\}$. For each question q_i, we tell the AI evaluator to prioritize responses from evaluated models instead of scoring them. We infer every pair of candidates to the AI evaluator swapping answer position (temperature $\tau = 0.01$). See Supp. Table 2(b) for the result of win-rate.

Report Generation. As shown in Table 2, we evaluate the quantitative result of automated CXR report generation. Our evaluation follows the official dataset split of MIMIC-CXR [10] and the dataset split with a proportion as [5,26] on IU-Xray [7]. We measured performance through metrics such as BLEU, METEOR, and ROUGE. WoLF outperformed the previous state-of-the-art method by a significant margin. Furthermore, as demonstrated in Fig. 5, our qualitative results validate that WoLF consistently generates content aligned with the Ground Truth (GT) report for various anatomical structures. See orange box in Supp. Fig. 2 for more qualitative report generation results. **Ablation and Other Studies.** In

Table 2. Report generation performance comparisons against other methods on MIMIC-CXR [10] and IU-Xray [7]. For a fair comparison with other methodologies, we directly quoted from the published literature.

Method	BLEU-1↑		BLEU-2↑		BLEU-3↑		BLEU-4↑		METEOR↑		ROUGE-L↑	
	MIMIC	IU	MIMIC	IU	MIMIC	IU	MIMIC	IU	MIMIC	IU	MIMIC	IU
R2Gen [5]	0.353	0.470	0.218	0.304	0.145	0.219	0.103	0.165	0.142	0.187	0.277	0.371
R2GenCMN [4]	0.353	0.475	0.218	0.309	0.148	0.222	0.106	0.170	0.142	0.191	0.278	0.375
M^2Tr [20]	0.378	0.486	0.232	0.317	0.154	0.232	0.107	0.173	0.145	0.192	0.272	0.390
CMCL [15]	0.344	0.473	0.217	0.305	0.140	0.217	0.097	0.162	0.133	0.186	0.281	0.378
CMN [3]	0.353	0.475	0.218	0.309	0.148	0.222	0.106	0.170	0.142	0.191	0.278	0.375
CA [17]	0.350	0.492	0.219	0.314	0.152	0.222	0.109	0.169	0.151	0.193	0.283	0.380
PPKED [16]	0.360	0.483	0.224	0.315	0.149	0.224	0.106	0.168	0.149	0.376	0.284	0.351
TransSQ [11]	0.423	0.484	0.261	0.333	0.171	0.238	0.116	0.175	0.168	0.207	0.286	0.415
MCGN [27]	0.373	-	0.235	-	0.162	-	0.120	-	0.143	-	0.282	-
ITA [25]	0.395	0.505	0.253	0.340	0.170	0.247	0.121	0.188	0.147	0.208	0.284	0.382
METransformer [26]	0.386	0.483	0.250	0.322	0.169	0.228	0.124	0.172	0.152	0.192	0.291	0.380
WoLF	**0.496**	**0.517**	**0.335**	**0.366**	**0.233**	**0.261**	**0.165**	**0.199**	**0.187**	**0.412**	**0.370**	**0.456**

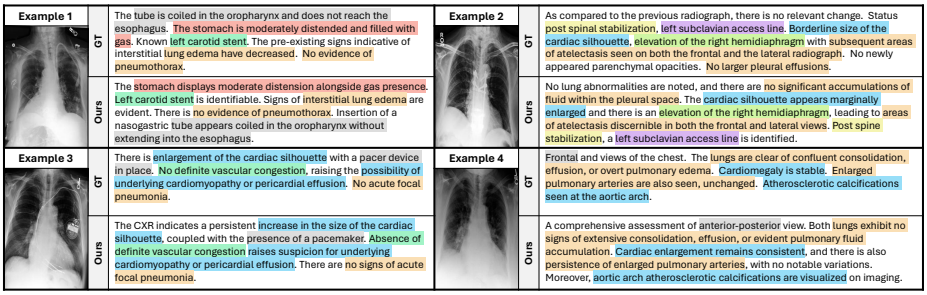


Fig. 5. Qualitative result of WoLF on report generation. Each highlighted color is mapped to the semantics of specific findings.

Supp. Table 2(a), we conduct ablation studies of Table 1. We examine our model learned without EHR and with EHR to the model during inference. Our ablation study For an additional comparison, please see Supp. Table 1(b), which shows traditional Visual Question Answering (VQA) experiments using the ELIXR [28] framework on the MIMIC-CXR dataset. In these experiments, we have achieved state-of-the-art results.

4 Conclusion and Future Work

This paper presents the first comprehensive exploration of LLM-based framework for understanding CXRs, encompassing data reformulation, model training, and evaluation strategies. We incorporate EHR into the model to enhance CXR understanding and generate instruction-following data reflecting real-world clinical processes. Furthermore, we refine CXR reports by decoupling them based on anatomical structures for training and introduce a novel masked attention mechanism to improve report generation performance. Additionally, our innovative AI evaluation protocol enables the assessment of LLMs from diverse perspectives. We

anticipate that our contributions, coupled with real-time EHR retrieval pipelines, will yield more adaptable frameworks for clinical decision-making.

References

1. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al.: Palm 2 technical report. arXiv preprint arXiv:2305.10403 (2023) 6, 7, 12
2. Bae, S., Kyung, D., Ryu, J., Cho, E., Lee, G., Kweon, S., Oh, J., Ji, L., Eric, I., Chang, C., et al.: Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. In: Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023) 4, 6, 7, 12
3. Chen, Z., Shen, Y., Song, Y., Wan, X.: 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). pp. 5904–5914 (2021) 8
4. Chen, Z., Shen, Y., Song, Y., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Aug 2021) 8
5. Chen, Z., Song, Y., Chang, T., Wan, X.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020 (2020) 7, 8
6. Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (2023), https://lmsys.org/blog/2023-03-30-vicuna/ 7
7. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016) 3, 7, 8
8. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2021) 7
9. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Moody, B., Gow, B., Lehman, L.w.H., et al.: Mimic-iv, a freely accessible electronic health record dataset. Scientific data **10**(1), 1 (2023) 4
10. Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019) 3, 4, 7, 8, 12
11. Kong, M., Huang, Z., Kuang, K., Zhu, Q., Wu, F.: Transq: Transformer-based semantic query for medical report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 610–620 (2022) 8
12. Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., Rastogi, A.: Rlaif: Scaling reinforcement learning from human feedback with ai feedback. arXiv preprint arXiv:2309.00267 (2023) 2, 6, 7
13. Lee, S., Kim, W.J., Ye, J.C.: Llm itself can read and generate cxr images. arXiv preprint arXiv:2305.11490 (2023) 1, 7, 12
14. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36** (2024) 4, 7, 12

15. Liu, F., Ge, S., Wu, X.: Competence-based multimodal curriculum learning for medical report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (2021) 8
16. Liu, F., Wu, X., Ge, S., Fan, W., Zou, Y.: Exploring and distilling posterior and prior knowledge for radiology report generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13753–13762 (2021) 8
17. Liu, F., Yin, C., Wu, X., Ge, S., Zhang, P., Sun, X.: Contrastive attention for automatic chest x-ray report generation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. pp. 269–280 (2021) 8
18. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024) 4
19. Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., Zhu, C.: G-eval: Nlg evaluation using gpt-4 with better human alignment. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 2511–2522 (2023) 2, 6
20. Nooralahzadeh, F., Gonzalez, N.P., Frauenfelder, T., Fujimoto, K., Krauthammer, M.: Progressive transformer-based generation of radiology reports. In: Findings of the Association for Computational Linguistics: EMNLP (2021) 8
21. OpenAI: Gpt-4 technical report (2023) 4
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021) 7
23. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 6, 7, 12
24. Thawkar, O., Shaker, A., Mullappilly, S.S., Cholakkal, H., Anwer, R.M., Khan, S., Laaksonen, J., Khan, F.S.: Xraygpt: Chest radiographs summarization using medical vision-language models. arXiv preprint arXiv:2306.07971 (2023) 1, 7, 12
25. Wang, L., Ning, M., Lu, D., Wei, D., Zheng, Y., Chen, J.: An inclusive task-aware framework for radiology report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 568–577 (2022) 5, 8
26. Wang, Z., Liu, L., Wang, L., Zhou, L.: Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11558–11567 (2023) 7, 8
27. Wang, Z., Tang, M., Wang, L., Li, X., Zhou, L.: A medical semantic-assisted transformer for radiographic report generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2022) 8
28. Xu, S., Yang, L., Kelly, C., Sieniek, M., Kohlberger, T., Ma, M., Weng, W.H., Kiraly, A., Kazemzadeh, S., Melamed, Z., et al.: Elixr: Towards a general purpose x-ray artificial intelligence system through alignment of large language models and radiology vision encoders. arXiv preprint arXiv:2308.01317 (2023) 8, 12
29. Yang, L., Wang, Z., Zhou, L.: Medxchat: Bridging cxr modalities with a unified multimodal large model. arXiv preprint arXiv:2312.02233 (2023) 12
30. Zhang, Y., Wang, X., Xu, Z., Yu, Q., Yuille, A., Xu, D.: When radiology report generation meets knowledge graph (2020) 5

— Supplementary Material —
WoLF: Wide-scope Large Language Model Framework for Chest X-ray Understanding

Seil Kang[1], Donghyun Kim[1], Junhyeok Kim[1], Hyo Kyung Lee[2], Seong Jae Hwang[1]

[1]Yonsei University, [2]Korea University
{seil, danny0103, timespt, seongjae}@yonsei.ac.kr,
hyokyunglee@korea.ac.kr

Table 1. (a) Ablation study of main paper Table 1. $^{-}$ for a given set of WoLF trained without EHRXQA [2]. $^{+}$ for a given set of WoLF given corresponding patient EHR while inference. **(b)** Win-rate results for each models. For instance, a $>$ **50%** win-rate signifies *left model* beats *right model* in more than a half of the questions.

Method	Gemini [23]					
	Accuracy↑	**Helpfulness↑**	**Relevance↑**	**Hallucination↑**	**Universality↑**	**Mean↑**
WoLF^{-}	71.58(±0.77)	72.10(±0.51)	69.27(±1.09)	78.14(±2.13)	74.96(±0.77)	73.21(±0.12)
WoLF	73.03(±0.64)	73.19(±0.87)	72.34(±0.42)	79.02(±0.67)	75.95(±0.49)	74.70(±0.34)
WoLF^{+}	**77.80**(±0.24)	**74.55**(±0.53)	**73.03**(±0.63)	**82.49**(±0.83)	**77.73**(±0.43)	**77.12**(±0.16)
	PaLM-2 [1]					
	Accuracy	**Helpfulness**	**Relevance**	**Hallucination**	**Universality**	**Mean**
WoLF^{-}	77.51(±0.98)	76.77(±1.36)	74.42(±2.75)	81.93(±1.19)	77.45(±2.60)	77.61(±1.26)
WoLF	79.54(±0.77)	79.93(±1.22)	**79.21**(±0.58)	85.30(±3.74)	78.18(±0.91)	80.13(±1.17)
WoLF^{+}	**81.09**(±0.30)	**81.60**(±1.56)	79.20(±0.55)	**86.22**(±3.62)	**78.84**(±1.89)	**81.39**(±1.65)

(a)

Comparison	**Accuracy↑**		**Helpfulness↑**		**Relevance↑**		**Hallucination↑**		**Universality↑**	
left model vs. *right model*	Gemini	PaLM-2	Gemini	PaLM-2	Gemini	PaLM-2	Gemini	PaLM-2	Gemini	PaLM-2
WoLF vs. LLaVA-Med [13]	71.50%	72.13%	77.75%	77.90%	76.25%	76.78%	84.25%	84.85%	73.40%	73.70%
WoLF vs. LLM-CXR [14]	64.05%	61.50%	61.73%	61.58%	61.95%	61.58%	63.93%	64.18%	63.6%	63.45%
WoLF vs. XrayGPT [24]	75.80%	75.83%	83.68%	83.15%	81.05%	82.05%	84.23%	84.68%	84.93%	84.63%
WoLF vs. **WoLF**$^{-}$	56.40%	57.10%	53.80%	54.50%	53.81%	54.00%	68.73%	68.80%	55.20%	55.40%
WoLF vs.**WoLF**$^{+}$	47.20%	47.45%	46.70%	47.12%	47.05%	47.95%	49.90%	49.95%	48.80%	49.20%

(b)

Table 2. Accuracy of the CXR-VQA task by topic. We utilize the ELIXR [28] framework for assessing this VQA performance in MIMIC-CXR [10].

Accuracy↑	**All↑**	**Presence↑**	**Location↑**	**Size, severity, type↑**
ELIXR [28]	54.8%	64.5%	41.0%	25.0%
XrayGPT [24]	25.2%	27.4%	21.9%	20.3%
LLM-CXR [13]	56.7%	60.1%	49.0%	53.1%
MedXChat [29]	61.2%	61.5%	56.3%	68.8%
WoLF	**62.1%**	**67.9%**	**56.9%**	**71.8%**

You are a helpful and precise assistant for checking the quality of the answer. We would like to request your feedback on the performance of three AI assistants in response to the user question displayed above. The user asks the question on an image. For your reference, the visual content in the image is represented with Visual Question Answering conversations of single turn between User and Assistant about the same image. The criteria for determining rates should follow the instructions below. \n

Evaluation Criteria: \n **Accuracy**: Use the single turn conversations as conversations to review both answers for accuracy. If any of the answers are inconsistent with what the reference says, consider the answer a bad answer. \n **Helpfulness**: A helpful answer is one that answers the question without evading what it asks. Also, a good answer is one that explicitly says yes or no to a question that can be answered with yes or no. \n **Relevance**: For an answer to be relevant, the content of the answer must be one of the things the question asks about. Even if the answer is long, but contains information that are not asked for in the question, consider it a bad answer. \n **Hallucination**: This item is for evaluating the AI assistant's hallucination. If an answer contains a content that conflicts with the source or cannot be verified by the factual knowledge, then consider the answer to be poor. \n **Universality**: Make an assessment of which answer is universally preferred. \nEach assistant receives an overall score on a scale of 0 to 10, where a higher score indicates better overall performance. Specifically, following the evaluation steps below, you should evaluate each assistant as objective as you can. \n

Scoring Steps: \n
1. First read the user question carefully and identify what the question is asking. \n **2**. Then using the evaluation criteria, evaluate the following answers from each assistant with comprehensive explanation of your evaluation. \n **3**. You have to rating each assistant's performance based on the evaluation criteria. The rating scale is from 0 to 10, where 1 is the worst and 10 is the best. \n **4**. Finally, referring to what you've answered, output lines containing only two values indicating the individual criteria scores and total scores for Assistant 1, 2, ... respectively. The two scores must be separated by a space(' ') \n **5**. Your answer should be followed below format. \n (ex. +++ Accuracy: {{score1: 0 to 10}} {{score2: 0 to 10}} \n Helpfulness: {{score1: 0 to 10}} {{score2: 0 to 10}} \n Relevance: {{score1: 0 to 10}} {{score2: 0 to 10}} \n Hallucination: {{score1: 0 to 10}} {{score2: 0 to 10}} \n Universal: {{score1: 0 to 10}} {{score2: 0 to 10}} +++ \n Evaluation Criteria Breakdown: ... \n A general description of the results: ...) **6**. Note that you should avoid any positional bias and ensure that the order in which the responses were presented does not affect your judgement.

Prioritizing Steps: \n
1. First read the user question carefully and identify what the question is asking. \n **2.** Then using the evaluation criteria, evaluate the following answers from each assistant. Provide a comprehensive explanation of your evaluation. \n **3.** Finally, referring to what you've answered, output a new single line containing only order for the answers, respectively. \n **4.** The numbers of ranking must be separated by a space and must not same number. For example, if the answers were 1st, 2nd best answers, then output must be 1 2. \n **5.** Your answer should be followed below format. \n (ex. +++ Ranking Accuracy: 1 2 \n Helpfulness: 1 2 \n Relevance: 1 2 \n Hallucination: 1 2 \n Universality: 1 2 +++ \n Given two answers, Assistant 1's response is better than Assistant 2's response.) \n **6.** Note that you should avoid any positional bias and ensure that the order in which the responses were presented does not affect your judgement.

Fig. 1. Evaluation prompts for scoring *(left)* and prioritizing *(right)*. Preamble and evaluation criteria are common in both scoring and prioritizing prompts. The two prompts, each for AI-evaluation *(left)* and for its ablation on Win-rate *(right)*, differ only in evaluation steps.

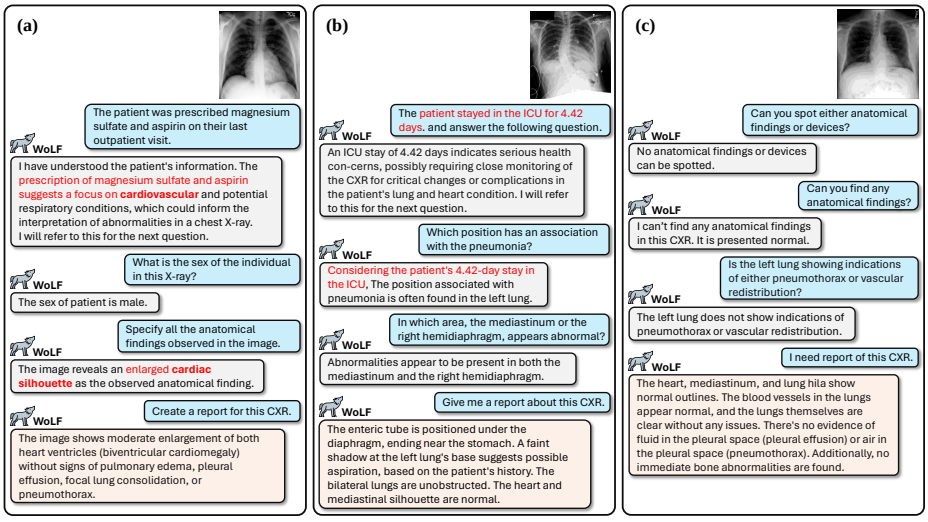


Fig. 2. Qualitative results of visual question-answering scenarios. As shown in (a) and (b), the model can be fed with patient histories and medication details from EHRs. WoLF utilizes these external contexts to deliver more accurate responses. (c) shows question-answering when there are no findings. The model answered correctly that no disease could be found, without causing any hallucinations.