

XrayGPT: Chest Radiographs Summarization using Large Medical Vision-Language Models

Omkar Thawakar*¹ Abdelrahman Shaker*¹ Sahal Shaji Mullappilly*¹ Hisham Cholakkal¹
 Rao Muhammad Anwer^{1,2} Salman Khan¹ Jorma Laaksonen² Fahad Shahbaz Khan¹

¹Mohamed bin Zayed University of AI ²Aalto University

Abstract

The latest breakthroughs in large vision-language models, such as Bard and GPT-4, have showcased extraordinary abilities in performing a wide range of tasks. Such models are trained on massive datasets comprising billions of public image-text pairs with diverse tasks. However, their performance on task-specific domains, such as radiology, is still under-investigated and potentially limited due to a lack of sophistication in understanding biomedical images. On the other hand, conversational medical models have exhibited remarkable success but have mainly focused on text-based analysis. In this paper, we introduce XrayGPT, a novel conversational medical vision-language model that can analyze and answer open-ended questions about chest radiographs. Specifically, we align both medical visual encoder (MedClip) with a fine-tuned large language model (Vicuna), using a simple linear transformation. This alignment enables our model to possess exceptional visual conversation abilities, grounded in a deep understanding of radiographs and medical domain knowledge. To enhance the performance of LLMs in the medical context, we generate 217k interactive and high-quality summaries from free-text radiology reports. These summaries serve to enhance the performance of LLMs through the fine-tuning process. Our approach opens up new avenues the research for advancing the automated analysis of chest radiographs. Our open-source demos, models, and instruction sets are available at: <https://github.com/mbzuai-oryx/XrayGPT>

1 Introduction

The Large-scale Vision-Language models have emerged as a transformative area of research at the intersection of computer vision and natural language processing, enabling machines to understand and generate information from both visual

and textual modalities. These models represent a significant advancement in the field, bridging the gap between visual perception and language comprehension, and have demonstrated remarkable capabilities across various tasks, including but not limited to image captioning (Hossain et al., 2019), visual question answering (Lu et al., 2023), and visual commonsense reasoning (Zellers et al., 2019). Training these models requires vast amounts of image and text data, enabling them to learn rich representations that capture the intricacies of both modalities. Additionally, fine-tuning can be employed using task-specific data to better align the models with specific end tasks and user preferences. Recently, Bard and GPT-4 have demonstrated impressive capabilities in various tasks, raising excitement within the research community and industry. However, it is important to note that the models of Bard and GPT-4 are not currently available as open-source, limiting access to their underlying architecture and implementation details.

Recently, Mini-GPT (Zhu et al., 2023) demonstrates a range of impressive capabilities by aligning both vision and language models. It excels at generating contextual descriptions based on the given image. However, it is not as effective in medical scenarios due to the significant differences between medical image-text pairs and general web content. Adopting vision-text pre-training in the medical domain is a challenging task because of two factors: (1) Lack of data, Mini-GPT has trained the projection layer on a dataset of 5M image-text pairs, while the total number of publicly available medical images and reports is orders of magnitude below. (2) Different modalities and domains, while Mini-GPT may involve distinguishing between broad categories like "Person" and "Car" the distinctions within medical domains are much more subtle and fine-grained. For instance, differentiating between terms like "Pneumonia" and "Pleural Effusion" requires more precision by capturing and

*First Author Equal Contribution

aligning crucial medical domain knowledge.

Chest radiographs are vital for clinical decision-making as they offer essential diagnostic and prognostic insights about the health of the patients. Text summarization tasks can partially address this challenge by providing meaningful information and summaries based on the given radiology reports. In our approach, we go beyond traditional summarization techniques by providing concise summaries that highlight the key findings and the overall impression based on the given X-ray. Additionally, our model allows for interactive engagement, enabling users to ask follow-up questions based on the provided answers. In this study, we stimulate the research around the automated analysis of chest radiographs based on X-ray images. Also, we argue that based on the visual and large language models, the majority of knowledge acquired during the pertaining stage of these models requires a domain-specific high-quality instruction set derived from task-specific data to achieve promising results. The main contributions of our work are:-

- The LLM (Vicuna) is fine-tuned on medical data (100k real conversations between patients and doctors) and 20k radiology conversations to acquire domain-specific and relevant features.
- We generate interactive and clean summaries (217k) from free-text radiology reports of two datasets: MIMIC-CXR (Johnson et al., 2019) and OpenI (Demner-Fushman et al., 2015). These summaries serve to enhance the performance of LLMs by fine-tuning the linear transformation layer on high-quality data.
- We align the frozen specialized medical visual encoder (MedClip) with a fine-tuned LLM (Vicuna), using a simple linear transformation to understand medical meanings and acquire remarkable visual conversation abilities.
- To advance the research in biomedical multimodal learning, we open-source our assets to the community: The codebase, the fine-tuned models, the high-quality instruction-set, and the raining recipe for data generation and model training are publically released.

2 Related Work

Medical Chatbot Medical chatbots have emerged as valuable tools in healthcare, providing personalized support, information, and assistance to pa-

tients and healthcare professionals. The recently introduced Chatdoctor (Li et al., 2023) is a next-generation AI doctor model that is based on the LLaMA (Touvron et al., 2023) model. The goal of this project is to provide patients with an intelligent and reliable healthcare companion that can answer their medical queries and provide them with personalized medical advice. After success of ChatGPT (OpenAI, 2022), GPT-4 (OpenAI, 2023) and other open source LLM’s (Touvron et al., 2023; Chiang et al., 2023; Taori et al., 2023), many medical chatbots were introduced recently such as Med-Alpaca (Han et al., 2023), PMC-LLaMA (Wu et al., 2023), and DoctorGLM (Xiong et al., 2023). These models utilize open source LLM’s and finetuned on specific medical instructions. These studies highlight the potential of medical chatbots to improve patient engagement and health outcomes through interactive and personalized conversations. Overall, medical chatbots offer promising opportunities using only textual modality for enhancing AI healthcare, with ongoing research focused on refining their capabilities with ethical considerations.

Large Language Vision Models A significant area of research in natural language processing (NLP) and computer vision is the exploration of Large Language-Vision Model (LLVM) learning techniques. This LLVM aims to bridge the gap between visual and textual information, enabling machines to understand and generate content that combines both modalities. Recent studies have demonstrated the potential of LLVM models in various tasks, such as image captioning (Zhu et al., 2023), visual question answering (Bazi et al., 2023; Liu et al., 2023; Muhammad Maaz and Khan, 2023), and image generation (Zhang and Agrawala, 2023).

3 Method

XrayGPT is an innovative conversational medical vision-language model specifically developed for analyzing chest radiographs. The core concept revolves around aligning medical visual and textual representations to enable the generation of meaningful conversations about these radiographs using our generated high-quality data. Our approach draws inspiration from the design of vision-language models in general, but with a specific focus on the medical domain. Due to the limited availability of medical image-summary pairs, we adopt a similar methodology by building upon a pre-trained medical vision encoder (VLM) and

medical large language model (LLM), as our foundation. The fine-tuning process involves aligning both modalities using high-quality image-summary pairs through a simple transformation layer. This alignment enables XrayGPT to possess the capability of generating insightful conversations about chest radiographs, providing valuable insights for medical professionals. By leveraging pre-existing resources and fine-tuning specific components, we optimize the model’s performance while minimizing the need for extensive training on scarce data.

3.1 Model Architecture

We show in Fig. 1 an overview of our XrayGPT. Given the X-ray, we align both visual features and textual information from a pre-trained medical vision encoder (VLM), and medical large language model (LLM). Specifically, we utilize Med-Clip (Wang et al., 2022) as a visual encoder and our large language model (LLM) is built upon the recent Vicuna (Chiang et al., 2023).

Given X-ray $\mathbf{x} \in R^{H \times W \times C}$, the visual encoder is used to encode the image into embeddings using a vision encoder E_{img} . Then, the raw embeddings are mapped to an output dimension of 512 using a linear projection head.

$$\mathbf{V}_p = f_v(E_{img}(\mathbf{x})) \quad (1)$$

where E_{img} is the vision encoder, f_v is the projection head.

To bridge the gap between image-level features and the language decoder’s embedding space, we employ a trainable linear transformation layer, denoted as t . This layer projects the image-level features, represented by \mathbf{V}_p , into corresponding language embedding tokens, denoted as \mathbf{L}_v :

$$\mathbf{L}_v = t(\mathbf{V}_p), \quad (2)$$

We have two text queries in our overall architecture. The first query, denoted as `###Assistant`, serves the purpose of determining the system role, which in our case is defined as *"You are a helpful healthcare virtual assistant."* The second text query, `###Doctor`, corresponds to the prompt itself. To ensure consistency, both queries undergo tokenization, resulting in dimensions represented by \mathbf{L}_t . Finally, \mathbf{L}_v is concatenated with \mathbf{L}_t and fed into the medical LLM, Fine-tuned Vicuna, which generates the summary of the chest x-ray.

Our XrayGPT follows a two-stage training approach. In the first stage, we train the model using

interactive summaries from MIMIC-CXR (Johnson et al., 2019) reports. While in the second stage, we use the high-quality curated interactive summaries of OpenI (Demner-Fushman et al., 2015) reports. MIMIC-CXR report findings contain information about patient history, which adds noise to the data. To mitigate this effect, we use small yet effective Interactive OpenI report summaries in the second stage to make our model robust to noise.

3.2 Image-text alignment

To align the generated high-quality summaries with the given x-ray, we use similar conversational format of the Vicuna (Chiang et al., 2023) language model as follows:

`###Doctor: $X_R X_Q$ ###Assistant: X_S`

where X_R is the visual representation produced by the linear transformation layer for image X , X_Q is a sampled question (e.g. What are the main findings and impression of the given X-ray?), and X_S is the associated summary for image X . In this way, we curate image-text pairs with detailed and informative interactive summaries.

4 Curating high-quality data

Datasets: The MIMIC-CXR consists of a collection of chest radiographs associated with free-text radiology reports. It consists of 377,110 images and 227,827 associated reports, which are used for both training and testing purposes. The dataset is de-identified by removing the health information to satisfy health insurance and privacy requirements. The OpenI dataset is a collection of chest X-ray images from the Indiana University hospital network, composing 6,459 images and 3,955 reports.

High Quality and Interactive Summaries: To generate concise and coherent medical summaries from the unstructured reports, we perform the following pre-processing steps for both datasets: (1) Removal of incomplete reports lacking finding or impression sections. (2) Elimination of reports that have finding sections containing less than 10 words. (3) Exclusion of reports with impression sections containing less than 2 words.

In addition, utilizing the power of gpt-3.5-turbo model, we further implement the following pre-processing techniques to ensure high-quality summaries per image: (1) Elimination of sentences containing comparisons to the patient’s prior medical history. (2) Removal of de-defined symbols `"_"`, while preserving the original meaning. (3)

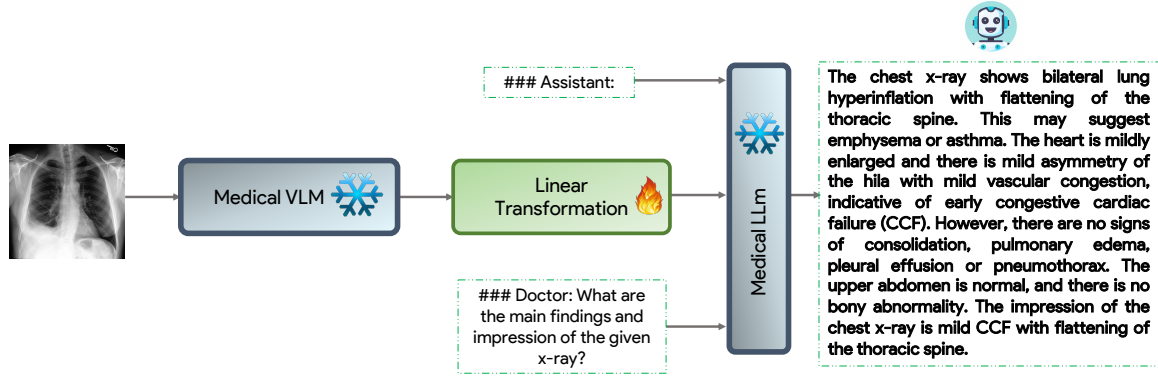


Figure 1: Overview of our XrayGPT framework. The input X-ray is passed sequentially to three components. (1) Frozen medical visual encoder to extract relevant features pertaining to the chest diagnosis. (2) Learnable linear transformation layer to align the medical visual features with the Medical LLM together to learn extensive medical visual-text alignment. (3) Frozen Medical LLM to generate a detailed summary of the X-ray based on the encoded features and the given prompt.

As our training relies on image-text pairs, we excluded the provided view from the summary. (4) We combine the clean findings and impressions to generate an interactive and high-quality summary.

Following these steps, we obtained a set of filtered training reports consisting of 114,690 reports associated with 241k training images based on Mimic-CXR dataset. Also, we obtained 3,403 high-quality summaries that used for training based on the OpenI dataset.

Here is an example before and after the proposed pre-processing. **Input findings:** *PA and lateral views of the chest were provided demonstrating no focal consolidation, effusion, or pneumothorax. Cardiomeastinal silhouette appears normal and stable. There is a compression deformity involving a mid thoracic vertebral body, which appears new from the prior chest radiograph of _____. No free air below the right hemidiaphragm. There are tiny surgical clips in the left base of neck, likely indicating prior thyroid surgery.* **Input Impression:** *No acute intrathoracic process. Interval development of a mid thoracic spine compression fracture.* **High-quality and interactive summary:** *The chest x-ray findings reveal no evidence of focal consolidation, effusion, or pneumothorax. The cardiomeastinal silhouette appears stable and normal. There is a newly developed mid thoracic spine compression fracture but no free air below the right hemidiaphragm. The presence of surgical clips in the left base of the neck suggests prior thyroid surgery. The impression suggests that there is no acute intrathoracic condition detected in the x-ray aside from the new development of mid thoracic spine compression fracture.*

5 Experiments

5.1 Implementation Details

Stage-1 Training: In stage-1 training, the model is designed to gain understanding of how Xray image features and corresponding reports are interconnected by analysing a large set of image-text pairs. The result obtained from the injected projection layer is considered as a gentle cue for our medically tuned LLM model, guiding it to produce the appropriate report based on the finding and impression that match the given x-ray images. We use high quality interactive report summary as described in sec. 4 of MIMIC-CXR (Johnson et al., 2019) train set with 213,514 image text pairs for training. During training the model trained for 320k total training steps with a total batch size of 128 using 4 AMD MI250X (128GB) GPUS.

Stage-2 Training: In stage-2 training, the pre-trained stage-1 model is enforced to gain radiology specific summary of how xray image features by examining set of highly curated image-text pairs from OpenI dataset. As a result, our medically tuned LLM can produce more natural and high quality radiology specific responses of given chest Xray images. We use high quality interactive report summary as described in sec. 4 from OpenI (Demner-Fushman et al., 2015) set with 3k image text pairs for training. During training the model trained for 5k total training steps with a total batch size of 32 using single AMD MI250X (128GB) GPU.

In both stage-1 and stage-2 of training, we utilize predetermined prompts in the given format:

###Doctor: <ImageFeature> <Instruction> ###Assistant:

Here, <Instruction> refers to a randomly selected instruction from our pre-established set of instructions, which includes different forms of instructions like "Describe the given chest x-ray image in detail." or "Are there any potential complications or risks associated with the observed abnormalities in this chest x-ray image? or the x-ray is normal." or "Is the overall impression provided by this chest x-ray image normal or abnormal? Answer based on the observed findings.". Similar to the baseline (Zhu et al., 2023), we do not compute the regression loss for this particular text-image prompt.

5.2 Evaluation Metrics

We used the Rogue Score as an evaluation metric to compare the contribution of our components over the baseline (Zhu et al., 2023). Rogue score has been commonly used (Cheng and Lapata, 2016; Nallapati et al., 2017; Zhang et al., 2020) to assess the quality of generated text, particularly in the field of natural language processing and text generation. It measures the overlap between the generated text and a set of reference texts, typically generated by human experts. The Rogue Score calculates precision, recall, and F1-score, taking into account the presence and ordering of n-grams (contiguous sequences of words). A higher Rogue Score indicates a better alignment between the generated text and the reference texts, indicating a higher level of quality and coherence. The Rogue Score serves as a valuable quantitative measure to objectively compare and assess the performance of different text generation models and techniques.

GPT-based evaluation of LMM (Language Model Mediated) generated text refers to the use of GPT (Generative Pre-trained Transformer) models to assess the quality and coherence of text generated by LMM approaches. LMM combines the power of pre-trained language models, such as GPT, with explicit control mechanisms to guide text generation. GPT-based evaluation involves using a fine-tuned GPT model to generate a set of reference texts, which can then be compared with the LMM-generated texts. Metrics such as perplexity, BLEU score, or the recently proposed Self-BLEU score can be used to quantitatively evaluate the similarity between the reference and generated texts.

5.3 Quantitative Measures

In this section, we highlight a key contribution of our XrayGPT compared to our baseline (Zhu et al., 2023). We conduct quantitative evaluation using ad-

Method	R-1	R-2	R-L
Baseline	0.1313	0.0221	0.0879
+ MedCLIP	0.1517	0.0308	0.0973
+ MedVicuna	0.2099	0.0551	0.1284
+ RadVicuna	0.3213	0.0912	0.1997

Table 1: Comparison of our XrayGPT components with Baseline (Zhu et al., 2023) using Rogue score on MIMIC-CXR (Johnson et al., 2019) Test set. All results are reported by progressively adding our components in the baseline Minigpt-4 (Zhu et al., 2023). Our approach outperforms the recent Minigpt-4 with an absolute gain of 19% in terms of R-1 score. Best results are in bold.

vanced metrics such as Rogue score and GPT-based evaluation as described in sec. 5.2. Tab. 1 shows comparison of our key components when progressively integrated into our baseline (Zhu et al., 2023) frame. From Tab. 1 our XrayGPT (row 4) has a significant improvement of 19% over the state-of-the-art baseline (Zhu et al., 2023) on the MIMIC-CXR test set. Also, we did LLM’s based evaluation by asking ChatGPT model to choose "which response is closer to reference between baseline vs XrayGPT" where our model scored 82% compared to baseline 6% showing the superiority of our XrayGPT for radiology-specific summary.

5.4 Qualitative Examples

We performed an extensive evaluation of our xrayGPT on a various Chest Xray samples where our model not only provides detailed findings and impression for given chest xray but also shows the conversation capability like radiologist. Please refer Figure 2 to 5 for more details.

6 Conclusion

To conclude, we present XrayGPT, a novel conversational medical vision-language model that combines both modalities to analyze and answer questions about chest radiographs. By aligning these models and leveraging our proposed interactive summaries from free-text radiology reports, XrayGPT demonstrates exceptional visual conversation abilities grounded in a deep understanding of chest radiographs in terms of findings and impressions. We open-source demos, models, and instruction sets. This work opens up new possibilities for advancing the automated analysis of chest radiographs and enhances the performance of large language models in the medical context.

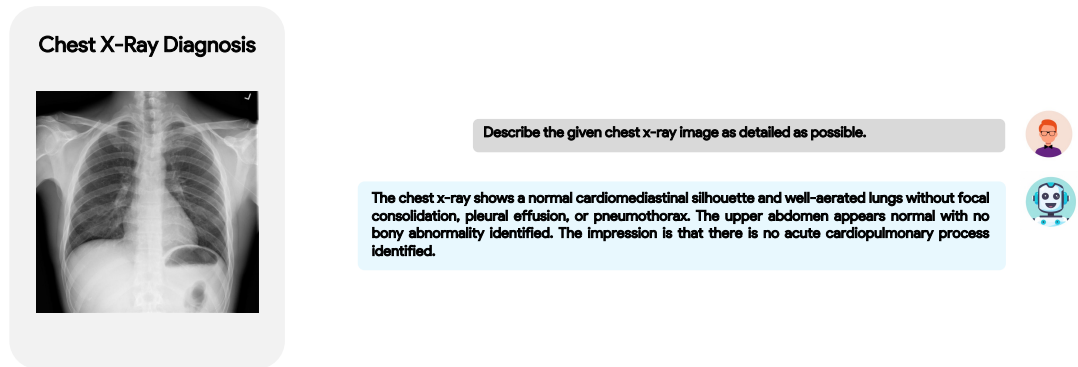


Figure 2: **XrayGPT: Delivering Comprehensive and Succinct Findings and Impressions for Every X-ray Image.** Our XrayGPT provides detailed radiology specific summary as well as concise findings and impression for given xray image.

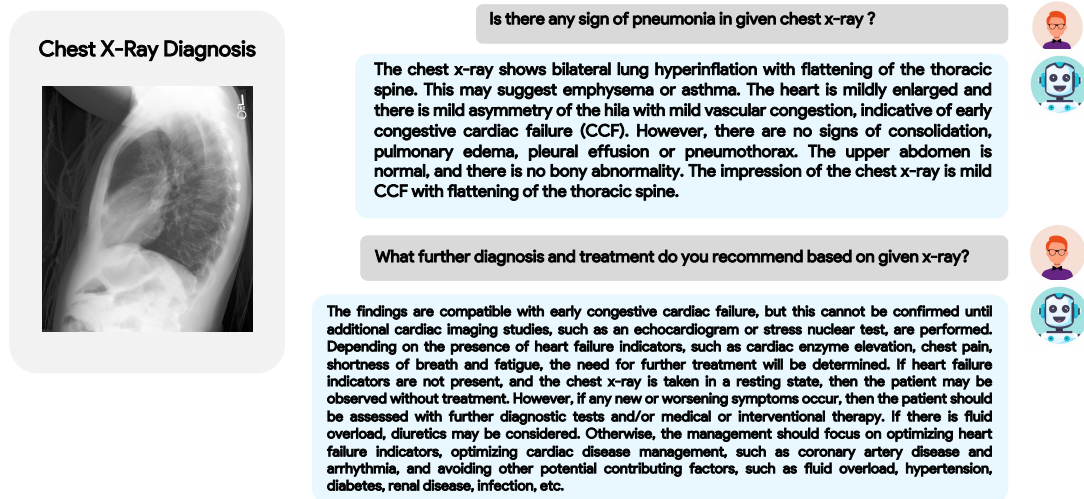


Figure 3: **XrayGPT: Empowering Users with Concise Insights - Explore the Comprehensive Analysis Derived from the xray Image.** Our XrayGPT provides a brief explanation for users queries based on the overall findings from the given xray image.

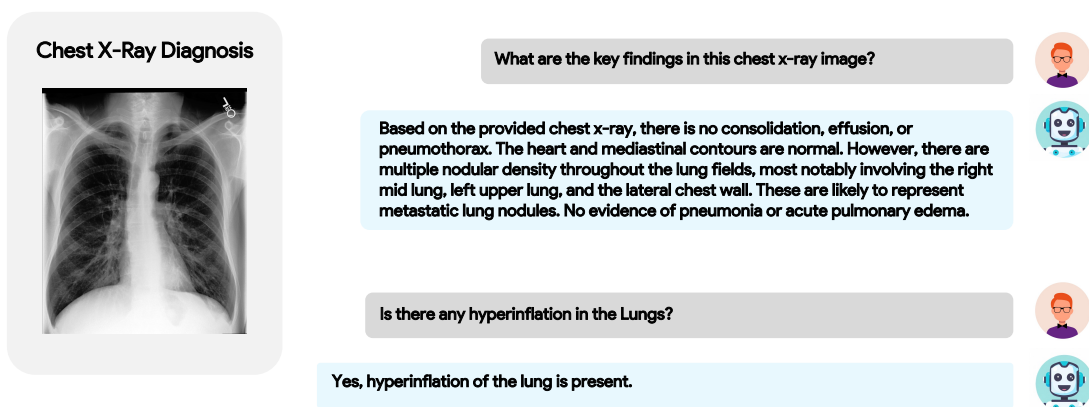


Figure 4: **XrayGPT: Unlocking the Power of Precision, Providing Specific Key Details in Response to User Queries.** Our XrayGPT is able to answer specific key detail in response to user queries.

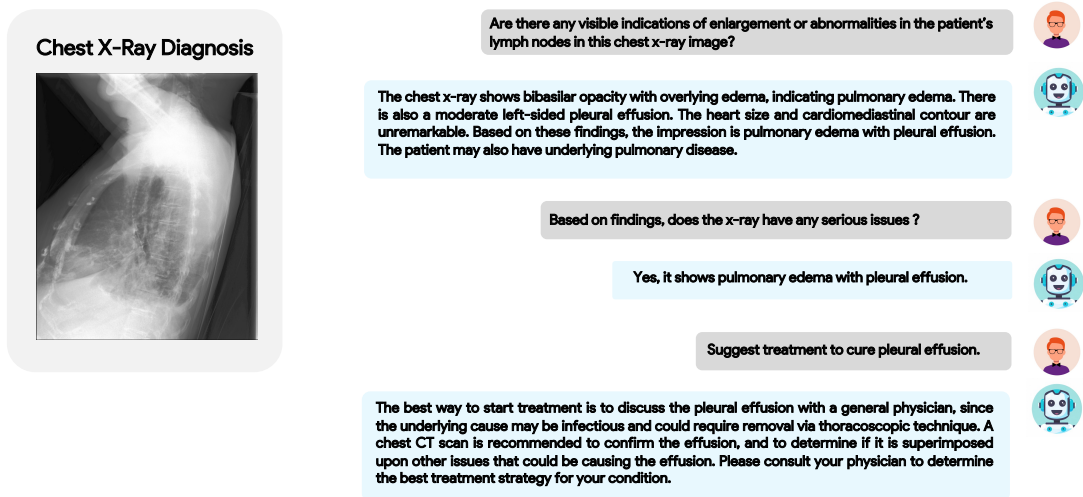


Figure 5: **XrayGPT: The Conversational AI Revolutionizing Radiological Interactions.** Our XrayGPT has radiological conversational capabilities.

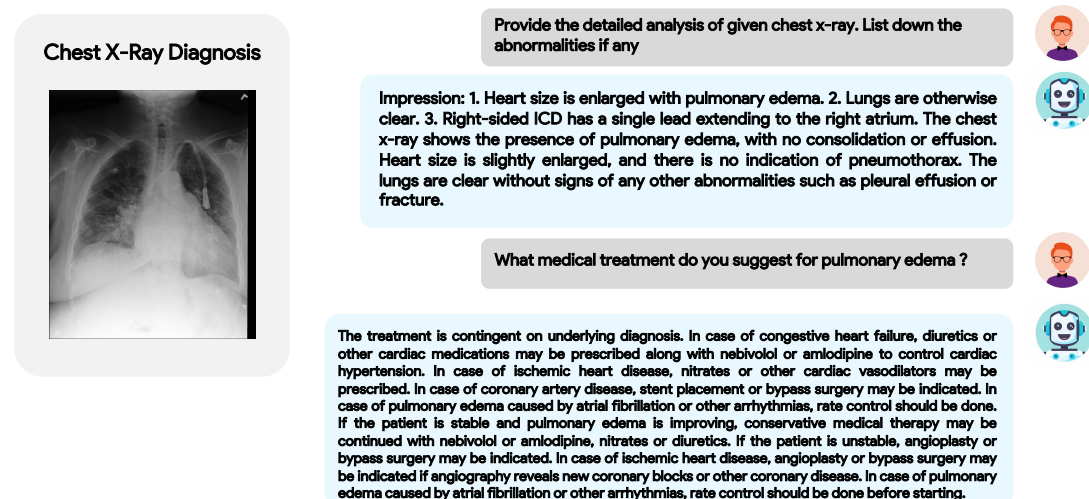


Figure 6: **XrayGPT: Medical Treatment Recommendation.** Our XrayGPT has the capability to suggest treatment based on the diagnosis.

References

- Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. Vision-language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380.
- Jianpeng Cheng and Mirella Lapata. 2016. Neural summarization by extracting sentences and words. *arXiv preprint arXiv:1603.07252*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. *Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality*.
- Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2015. *Preparing a collection of radiology examinations for distribution and retrieval*. *Journal of the American Medical Informatics Association*, 23(2):304–310.
- Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressen. 2023. Medalpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.
- MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36.
- Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. *Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports*. *Scientific Data*.
- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023. *Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge*.
- Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. 2023. Q2atransformer: Improving medical vqa via an answer querying decoder. *arXiv preprint arXiv:2304.01611*.
- Siyu Lu, Yueming Ding, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023. Multiscale feature extraction and fusion of image and text in vqa. *International Journal of Computational Intelligence Systems*, 16(1):54.
- Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*.
- Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.
- OpenAI. 2022. *Chatgpt*.
- OpenAI. 2023. *Gpt-4 technical report*.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. *Medclip: Contrastive learning from unpaired medical images and text*. *arXiv preprint arXiv:2210.10163*.
- Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further fine-tuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.
- Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.
- Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.
- Lvmin Zhang and Maneesh Agrawala. 2023. *Adding conditional control to text-to-image diffusion models*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. *Minigt-4: Enhancing vision-language understanding with advanced large language models*. *arXiv preprint arXiv:2304.10592*.