ChatXRay: A Multimodel Assistant for Chest Radiographs Summarization

Anonymous CVPR submission

Paper ID *****

Abstract

Recent advancements in large vision-language models, exemplified by LLaVA and GPT-4, have demonstrated remarkable capabilities across a broad spectrum of tasks. These models are trained on extensive datasets comprising billions of public image-text pairs, enabling them to excel in diverse tasks. Despite their rapid progress, these general-domain vision-language models still lack the sophistication required to comprehend and discuss biomedical images effectively. In response to this gap, we introduce ChatXRay, a pioneering multimodal model designed to function as an assistant for generating medical reports based on chest radiographs. We trained out model on puclic datasets, and our open-source codes and models weights are available at: https://github.com/Frankgu3528/ChatXRay.

1. Introduction

Recent year have witnessed the rapid development of large language models emerged as prominent tools in the realm of natural language processing. LLMs are trained on huge scale datasets and show promising abilities on several downstream tasks, including text generation, poblem solving, etc.

Many researchers have worked to align large language models with vision models, which extends the models' modalities and enable them to understand pictures. These large-scale vision-language models demonstrate promising potential in solving real-world vision-central problems. The most known is GPT4 [8], which is close source and powerful. Some open source VLLMs are miniGPT4 [13] and LLaVA [6], which also show significant abilities in understanding vision problems. There's also many VLLMs trained on large quantities on chinese corpus, including Visual-GLM [2] ,Qwen-VL [1].

Healthcare is a domain many scientists focus on. Local hospitals always face a problem that there's not enough professional doctors to serve for all the large amount of patients. So researchers have developed many medical chatbots to help address the issue. For example,ChatDoctor [5], Clinical Camel [10], DoctorGLM [12]. These use large language models (ChatGLM,Camel,LLama) as base model and collect large quantities of high quality medical conversations to make LLMs understand more about medical domain.

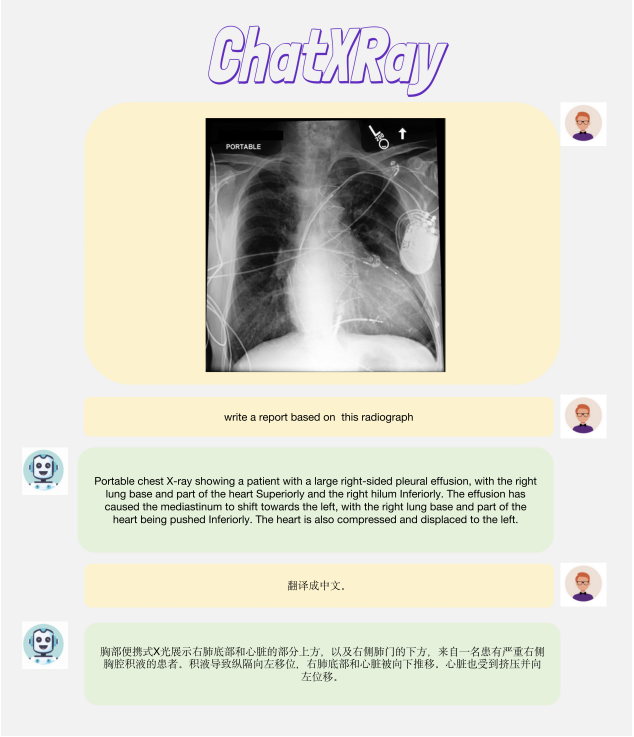

Figure 1. Demo of our ChatXRay

However, this chatbots only support text input because they are finetuned on large language models. So in this paper, we introduce ChatXRaya vision Large language model that helps patients use natural language to interact with radiographs. We believe ChatXRay can help patients preliminarily diagnose their conditions, thus improving the efficiency of hospital consultations. In summary, our paper makes the following contributions:

- We collect 85K chest radiographs and medical reports

pairs from public datasets. We finetuned a ultimodel large language model(Qwen-vl) on our dataset to acquire domain-specific abilities.

- We opensource all of our code, including data downloading, data processing, finetuning and comparison code with other baseline models.

2. Related Work

Biomedical Chatbots: The development of biomedical chatbots has been greatly influenced by the success of open-sourced instruction-tuned large language models (LLMs) such as ChatGPT in the general domain. Notable examples of these biomedical LLM chatbots include ChatDoctor [5], Med-Alpaca [3], PMC-LLaMA [11], Clinical Camel [10], DoctorGLM [12]. These chatbots have been designed with the specific aim of addressing various medical and healthcare-related queries and needs.

Multimodel Large Language Models After seeing the promising power of ChatGPT, researchers in computer vision and natural language processing are thrilled to build some multimodel large language models(MLLMs). They are pre-trained on large quantities of image-text paires . This MLLMS aims to bridge the gap between different modalities, enabling machines to understand and generate content that combines both textual and image informations. Some of them are LLaVA,

Multimodel Medical Chatbots In addition to the development of biomedical chatbots, there has been a notable surge in the creation of multimodal medical chatbots, specifically tailored to address the analysis and interpretation of radiographs. These advanced chatbots leverage a combination of textual and visual information to provide comprehensive insights into medical imaging data. LLava-Med [4] is based on LLava desighed by microsoft. XrayGPT [7] uses miniGPT4 as a base model. XrayGLM [9]is the first chinese version chatbots based on ChatGLM.

3. Method

ChatXRay is designed based on Qwen-VL [1]. Qwen-VL is a large-scale vision-language model (LVLMs) designed to perceive and understand both texts and images. I Large Language Model: Qwen-VL adopts a large language model as its foundation component. The model is initialized with pre-trained weights from Qwen-7B (Qwen, 2023).

Table 1: Details of Qwen-VL model parameters.

Vision Encoder	VL Adapter	LLM	Total
1.9B	0.08B	7.7B	9.6B

Visual Encoder: The visual encoder of Qwen-VL uses the Vision Transformer (ViT) (Dosovitskiy et al., 2021) architecture, initialized with pre-trained weights from Openclip's ViT-bigG (Ilharco et al., 2021). During both training and inference, input images are resized to a specific resolution. The visual encoder processes images by splitting them into patches with a stride of 14, generating a set of image features.

Position-aware Vision-Language Adapter: To alleviate the efficiency issues arising from long image feature sequences, Qwen-VL introduces a vision-language adapter that compresses the image features. This adapter comprises a single-layer cross-attention module initialized randomly.

3.1. Finetuning on Domain-Specific data

Qwen-VL already has great ability in understanding image in general domains, however, it still lacks knowledge in analysing Chest Radiographs. So we collect 80000+ radiographs from Mimic chest xray dataset, each with medical report written by professinal doctors attach to it.

We generally follows Qwen-VL's finetuning strategy. To make save gpu memory usage, we use lora strategy to finetune our model.

In the finetuning stage, the model was trained with lora on image-text pairs. We trained the model for 1 epoch with a total batch size of 128 using a single Nvidia A100 GPU for 28 hours. You can see the visualization of training loss and learning rate in Figure 3.

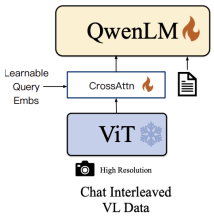

Figure 2. The original finetuning process

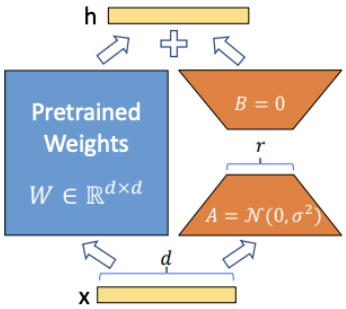

Figure 3. The architecture of Lora

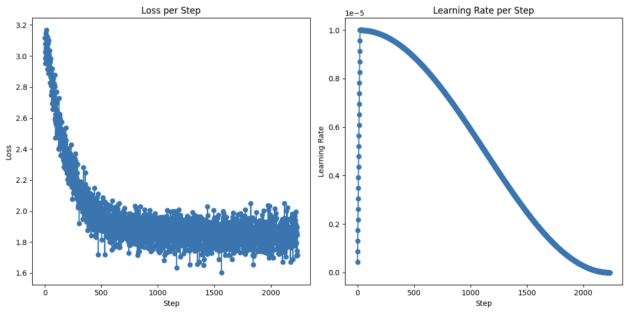

Figure 4. Visualization of the training loss and learning rate

4. Experiments

4.1. Datasets

We collect two high-quality open source datasets: Mimic-Chest-Xray(V1) and Openi. You can view an example at Figure 4.

Stable cardiac enlargement. Increasing perihilar vasculature and indistinctness favoring developing mild perihilar edema. No pleural effusions. No pneumothorax.

Endotracheal tube has its tip 4.5 cm above the carina. Nasogastric tube courses below the diaphragm with the tip not identified.

Figure 5. example of a radiograph-report pair in our dataset

4.2. Evaluation

4.2.1 Evaluation Metrics

To evaluate the performance of out ChatXRay model, we collect 100 Chest Radiographs and doctors' reports pairs. We ask the model to make a report to each graph. Than we ask GPT4 which model's report is more precise and acurate. Our prompt is like below:

```
messages=[
        {"role": "system", "content": "I will give you three reports.
The first is the true one, the next two are made by two models.
Please tell me which one can point out the symptom more
precisely, give me the result in fuyu or chatxray and don't
explain."},
        {"role": "user", "content": "Truth: "+entry["ground_truth"]},
        {"role": "user", "content": "fuyu: "+entry["fuyu-8b"]},
        {"role": "user", "content": "chatxray: "+entry["ChatXRay"]}
    ]
```

Figure 6. example prompt for evaluation

We then use our evaluation dataset(100 images and report pairs) to run the comparison between two models. If GPT4 thinks A is more precise in most times, we assert model A is better, otherwise we thinks B is better at giving reports.

4.2.2 Baselines

We entensively compare our model with 4 advanced methods, which can be divided to two categories, including 1) Universal Multimodel Large language models 2Multimodel models served as medical chatbots.

To be precise, we choose Fuyu-8b[1],a multi-modal model that can consume images and text and produce text, to represent VLLMs trained on general domains. Architecturally, Fuyu is a vanilla decoder-only transformer - there is no image encoder. Image patches are instead linearly projected into the first layer of the transformer, bypassing the embedding lookup.

4.2.3 Evaluation Results

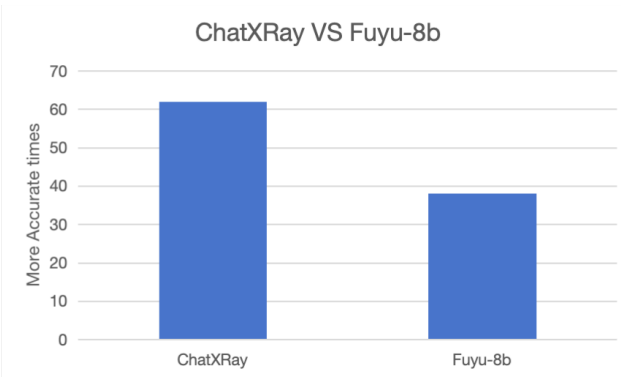

Figure 7. ChatXRay compared to Fuyu-8b

Ablation study Since ChatXRay is finetuned on its base model: Qwen-VL, we also do some ablation study to see whether our finetune process really make a difference in understanding radiographs. We ask ChatXRay and Qwen-VL to give report towards the same 100 radiographs, and follow the same procedure to test which one make better response each time. The results are as follows. We see ChatXRay performs slightly better than Qwen-VL.

4.3. Discussion

We'd like to mention that maybe because Qwen-VL's is pre-trained on large amount of Chinese corpus, when we tried to infer the model with English input, we sometimes see the output are mixed with Chinese characters.Here's two examples:

[1] https://www.adept.ai/blog/fuyu-8b

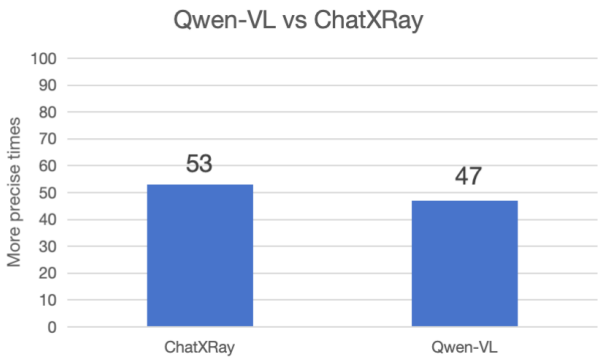

Figure 8. ChatXRay vs Qwen-VL

Also, we are doubting whether the finetuned model really learned how to understand chest radiographs and catch the detail symptoms to make currect diagnosis. As far as we have learned, our model and other model (XrayGLM [9])'s report accuracy are low. We may doubt that after finetuning, maybe the model only know how to talk like a real doctor. It's possible that model doesn't catch small details and only talk like doctors to make reports. And because many reports have a similar style, the results seem fine but are not meaningful. So we call on hospitals to open source more chest graphs to help construct better fine-tune dataset and we hope researchers to develop more powerful base models.

5. Conclusion

To conclude, we presented ChatXRay, a novel multimodel medical vision-language model that combines both modalities to analyze and give reports about chest radiographs. We collected large number of expert labeled chest radiographs from public datasets, and use lora techniques to funetune the powerful vision-language model: Qwen-VL. We conducted several experiments to show ChatXRay has learned medical after the finetune process and is better than giving report towards chast radiographs than vLLMs from general domain.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2

[2] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835, 2021. 1

[3] Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023. 2

[4] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day, 2023. 2

[5] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge, 2023. 1, 2

[6] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1

[7] Sahal Shaji Mullappilly Hisham Cholakkal Rao Muhammad Anwer Salman Khan Jorma Laaksonen Omkar Thawkar, Abdelrahman Shaker and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using large medical vision-language models. *arXiv: 2306.07971*, 2023. 2

[8] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul

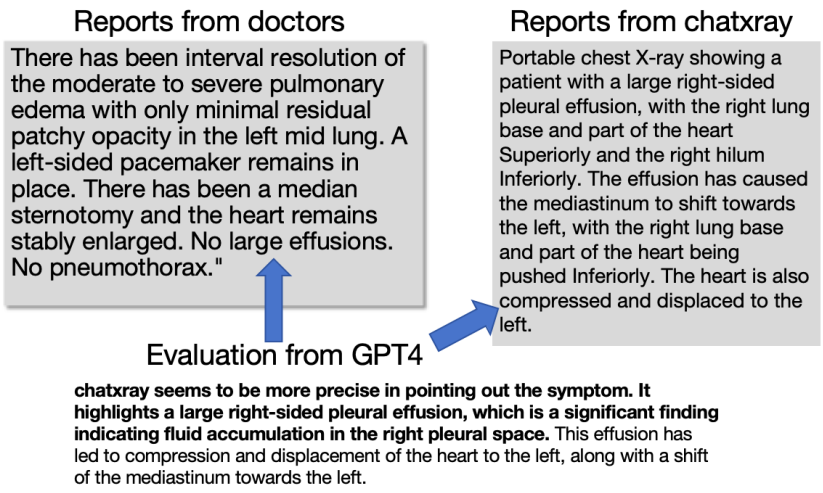

Figure 9. Example of our models output

This is a chest Xray image showing a right-sided pleural effusion with a rightward shift of the mediasternal lines and a decrease in the span of the chest. The effusion has caused a compression at the right lung bases, with a弓弦状阴影. The left lung is normal in size and contour. The heart and trachea are also normal in position and appearance. The image also shows a rightward shift of the trachea and a decrease in the span of the chest. The effusion has caused a compression at the right lung bases, with a弓弦状阴影. The left lung is normal in size and contour. The heart and trachea are also normal in position and appearance.

Figure 10. Example 1: Chinese character problem

The chest X-ray image shows a large right-sided pleural effusion with弓背屈位的右侧胸部。右侧肺野被积液广泛压缩，可见多个大小不等的液气泡。右侧心脏形态、位置正常，未见异常阴影。肺动脉段、右肺中肺动脉周围见多发性钙化灶。左侧胸部正常。因此，该患者的胸部X线片提示有右侧大量胸腔积液，且存在多发性钙化灶。

Figure 11. Example 2: Chinese character problem

McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. 1

[9] Junrong Li Patrick Pang Rongsheng Wang, Yaofei Duan and Tao Tan. Xrayglm: The first chinese medical multimodal model that chest radiographs summarization. https://github.com/WangRongsheng/XrayGLM, 2023. 2, 4

[10] Augustin Toma, Patrick R. Lawler, Jimmy Ba, Rahul G. Krishnan, Barry B. Rubin, and Bo Wang. Clinical camel: An open expert-level medical language model with dialogue-based knowledge encoding, 2023. 1, 2

[11] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-llama: Towards building open-source language models for medicine, 2023. 2

[12] Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Linlin Huang, Qian Wang, and Dinggang Shen. Doctorglm: Fine-tuning your chinese doctor is not a herculean task, 2023. 1, 2

[13] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language

understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 1