

Self-Training Large Language and Vision Assistant for Medical Question-Answering

Guohao Sun¹, Can Qin², Huazhu Fu³, Linwei Wang¹, Zhiqiang Tao¹,

¹Rochester Institute of Technology, ²Salesforce AI Research,

³Institute of High Performance, Computing, Agency for Science, Technology and Research
{gs4288, linwei.wang, zhiqiang.tao}@rit.edu
cqin@salesforce.com, hzfu@ieee.org

Abstract

Large Vision-Language Models (LVLMs) have shown significant potential in assisting medical diagnosis by leveraging extensive biomedical datasets. However, the advancement of medical image understanding and reasoning critically depends on building high-quality visual instruction data, which is costly and labor-intensive to obtain, particularly in the medical domain. To mitigate this data-starving issue, we introduce **Self-Training Large Language and Vision Assistant for Medicine (STLLaVA-Med)**. The proposed method is designed to train a policy model (an LVLM) capable of auto-generating medical visual instruction data to improve data efficiency, guided through Direct Preference Optimization (DPO). Specifically, a more powerful and larger LVLM (e.g., GPT-4o) is involved as a biomedical expert to oversee the DPO fine-tuning process on the auto-generated data, encouraging the policy model to align efficiently with human preferences. We validate the efficacy and data efficiency of STLLaVA-Med across three major medical Visual Question Answering (VQA) benchmarks, demonstrating competitive zero-shot performance with the utilization of only 9% of the medical data. Our implementation is available at <https://github.com/heliossun/STLLaVA-Med>.

1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated impressive performance across a wide range of medical challenges (Li et al., 2023; Moor et al., 2023; Hu et al., 2024) by fine-tuning through biomedical visual instruction data. Similar to general LVLMs (Liu et al., 2023a; Chen et al., 2023), existing methods tailored for biomedical tasks primarily focus on collecting high-quality medical data to enhance task generalization and visual understanding. However, collecting medical data necessitates specialized expertise from physicians and raises privacy concerns, making the process both time-consuming and costly. To address

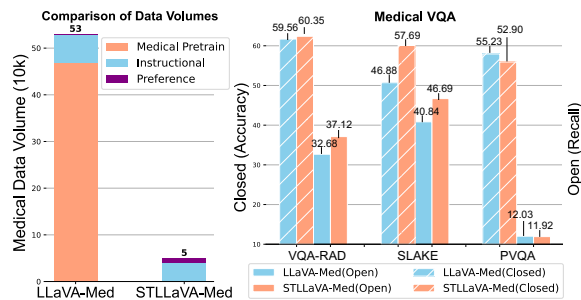


Figure 1: *Left*: Comparison of total medical data usage between LLaVA-Med (530K) and STLLaVA-Med (50k). *Right*: Comparison results on three medical VQA datasets. STLLaVA-Med reports better/comparable performance, using much less medical training data.

this data-starving issue, recent studies (Li et al., 2023) have explored leveraging larger models/APIs (e.g., GPT-4 (Achiam et al., 2023)) to generate medical data. Nevertheless, this kind of method does not fully resolve the high API costs (Deng et al., 2024) associated with building instructional data and still requires large-scale pre-training data to align medical images and text (see Fig. 1 left).

To bridge the gap in medical data acquisition, we propose **Self-Training Large Language and Vision Assistant for Medicine (STLLaVA-Med)**, a new training pipeline that enables LVLMs to automatically generate medical instruction data governed by Direct Preference Optimization (DPO) (Rafailov et al., 2023). Different from previous self-training approaches (Wang et al., 2023; Deng et al., 2024), which generate answers for fixed/pre-defined questions (e.g., summarization and report), this work automatically generates open-ended questions and answers them, to enhance the diversity of self-training data and further improve medical image reasoning.

Moreover, achieving precise control of the generated model response is also challenging due to its unsupervised nature (Rafailov et al., 2023; Zhao et al., 2023; Azar et al., 2023; Mehta et al., 2023). Existing methods for gaining such steerability, such as reinforcement learning (Ouyang et al., 2022) and

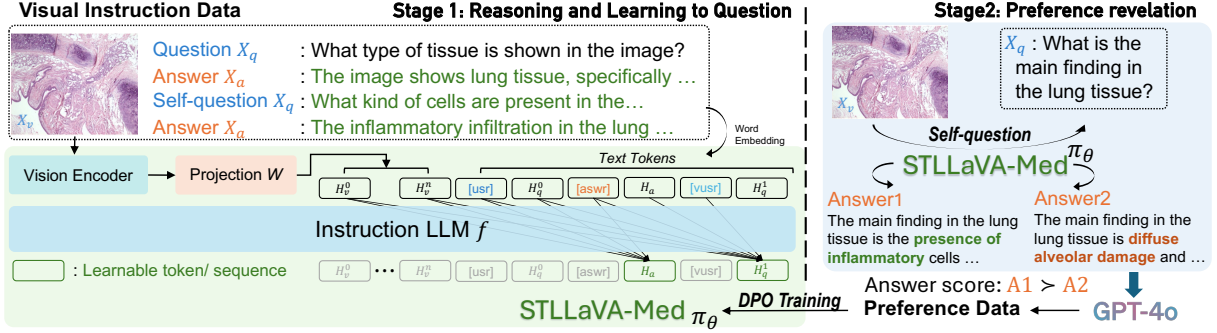


Figure 2: Model architecture of STLLaVA-Med and self-training pipeline. *Left*: stage 1 aiming to optimize the model π_θ improving medical image reasoning and learning to question. *Right*: in stage 2, we first prompt π_θ to auto-generate preference data under the guidance of GPT-4o, then supervise π_θ for DPO fine-tuning.

DPO (Rafailov et al., 2023) from human feedback, mainly rely on collecting human labels to evaluate the relative quality of model generations and fine-tune the unsupervised LVM to align with human preferences, which still burdens data collection in biomedical domains. To this end, the proposed STLLaVA-Med implements DPO by leveraging a larger LVM with better general medical knowledge to supervise the policy model.

Overall, the proposed STLLaVA-Med realizes self-training in two stages – 1) *reasoning and learning to question* and 2) *preference revelation*. In Stage 1, to enhance the model’s reasoning and questioning skills, we incorporate questions within the visual instructional data as an additional learning objective following (Sun et al., 2024b). After the first-stage training, STLLaVA-Med can generate question-answer pairs automatically. In Stage 2, we leverage GPT-4o (OpenAI, 2024) as a medical expert to further supervise fine-tuning STLLaVA-Med through DPO, ensuring it adheres to our designed preferences (e.g., detail, relevance, and accuracy) on the auto-generated data. We summarize the contributions of this work as follows:

- We propose a novel self-training approach for LVMs that enhances medical reasoning skills with less medical data. Our approach improves the data efficiency of training LVMs for specific domains.
- The proposed STLLaVA-Med enables the automatic construction of medical instructional data, supervised by a stronger and heavy LVM (i.e., GPT-4o) and governed through DPO, which allows our LVM to adhere to preferences in a self-training way.
- Experiments on three major medical VQA benchmarks demonstrate that our method

achieves highly competitive zero-shot performance compared to existing methods yet utilizing only 9% of the medical data.

2 STLLaVA-Med

In this section, we introduce STLLaVA-Med (see Fig. 2) given by our proposed two-stage self-training algorithm, which is designed to enhance the data efficiency when training an LVM for medical tasks. Specifically, we optimize the LVM – a policy model – in two stages sequentially. The policy model π_θ parameterized by θ first learns to automatically generate question-answer pairs for self-training, then utilizes DPO to control the prediction behavior precisely.

Stage1: Reasoning and learning to question.

The main part of self-training is automatic question generation and answering. Specifically, we follow (Sun et al., 2024b) by adding a special token $\langle vusr \rangle$ and set the question-tokens to learnable, to jointly fine-tune π_θ for reasoning and questioning on visual instructional data \mathcal{D}_{ft} .

Given the visual instruction data $\mathcal{D}_{ft} = \{(X_v, X_c)\}_1^N$, where the conversation $X_c = \{X_q^{(j)}, X_a^{(j)}\}_{j=1}^M$ consists of M QA pairs, the text X_c , and the image X_v are encoded to sequential embeddings as $H_c = \{H_q^{(j)}, H_a^{(j)}\}_{j=1}^M$ and H_v by word embedding and vision encoder. We minimize the negative log-likelihood loss for the vq : *visual questioning* and qa : *answering* as the following:

$$\mathcal{L}_{vq} = \sum_{v,c \in \mathcal{D}_{ft}} -\log \pi_\theta(H_q^{(j+1)} | H_v, H_c^{(1:j)}),$$

$$\mathcal{L}_{qa} = \sum_{v,c \in \mathcal{D}_{ft}} -\log \pi_\theta(H_a^{(j+1)} | H_v, H_c^{(1:j)}, H_q^{(j+1)}),$$

where $j \in \{1, \dots, M\}$ indicates the index of question or answer within the conversational data $H_c = \{H_q^{(j)}, H_a^{(j)}\}_{j=1}^M$. To mitigate the heavy computational overhead, we fine-tune LoRA (Hu

et al., 2022) in both the vision encoder and LLM. Thus, the learnable parameters θ of the policy model π_θ during fine-tuning represent a combination of all the parameters of LLM-LoRA, ViT-LoRA, and the vision-to-language projector. In addition, we skip the vision-language alignment on medical image-text pairs for data efficiency by loading the pre-trained weights (Sun et al., 2024b) fine-tuned on general-purpose visual instructional data. After training with these objective functions, our model can raise and answer questions for a medical image, enabling automatic QA generation in the following stage.

Stage2: Preference revelation.

We apply DPO (Rafailov et al., 2023) to fine-tune the unsupervised LVLM to align with pre-defined preferences. Unlike previous works (Rafailov et al., 2023; Zhao et al., 2023), we employ π_θ to automatically generate a preference dataset $D_{pref} = \{(X_v, X_q, X_{a_w}, X_{a_l})\}_1^N$. Specifically, we prompt π_θ to generate an image-related question X_q and two different answers X_a , which are labeled as *win* X_{a_w} and *loss* X_{a_l} answers by GPT-4o. For all the experiments, we first map X_q, X_{a_w}, X_{a_l} , and X_v to H_q, H_{a_w}, H_{a_l} , and H_v using the same word embedding and vision encoder in stage 1, and fine-tune π_θ through DPO by minimizing the following negative log-likelihood loss:

$$\mathcal{L}_{DPO}(\pi_\theta; \pi_{ref}) = -\mathbb{E}_{v,q,a_w,a_l \in \mathcal{D}_{pref}} [\log \sigma(\beta \log \frac{\pi_\theta(H_{a_w} | H_v, H_q)}{\pi_{ref}(H_{a_w} | H_v, H_q)}) - \beta \log \frac{\pi_\theta(H_{a_l} | H_v, H_q)}{\pi_{ref}(H_{a_l} | H_v, H_q)}), \quad (1)$$

where β is a parameter controlling the deviation from the base reference policy π_{ref} , which prevents the policy model from deviating too far from the distribution of correct generation, as well as maintaining the generation diversity and preventing mode-collapse to single high-reward answers (Rafailov et al., 2023). Notably, π_θ and π_{ref} are initialized from the same weights at beginning, and only π_θ is optimized during training. In this way, we fit an implicit reward to precisely control the model generation by the pre-defined preference such as accuracy and detail.

3 Self-training Datasets

Medical LVLMs (Li et al., 2023; Zhang et al., 2023; Moor et al., 2023) generally adopt pre-training on massive medical data, to realize medical image-text alignment. However, the proposed STLLaVA-Med does not involve such a medical corpus pre-training, providing new insights into data efficiency. To

Table 1: Statistics of medical training data.

Method	#Images	#QA-Pairs
LLaVA-Med _{pt}	467710	467710
LLaVA-Med _{ft}	56708	164231
ours	37452	108545

fine-tune the LVLM for medical tasks, we utilize a filtered open-source medical instructional dataset Med-60k-IM (Li et al., 2023) as \mathcal{D}_{ft} due to image unavailability. Table 1 provides the medical data statistics for training LLaVA-Med and STLLaVA-Med. We show how to employ the policy model to auto-generate a preference dataset for DPO fine-tuning in the following process:

Auto-generated Questions. We randomly sample 10k medical images from Med-60k-IM datasets and prompt π_θ to generate questions.

GPT-4o guided preference data collection. We prompt π_θ to predict two answers to each generated question. Specifically, to ensure the difference between answers, we set the temperature scaling to 1.2, $TopK = 100$, and $TopP = 0.95$, encouraging the model to generate more diverse and non-repetitive output. In previous research (Rafailov et al., 2023), the preference data were annotated by human annotators. In contrast, this work utilizes GPT-4o as a simulated expert since we observe its excellent biomedical performance (Yue et al., 2023) and the best downstream task performance in Table 2. We prompt GPT-4o (see Appendix 4.2 for prompt design) with all the information to label the answers with *win* or *loss*, treated as X_{a_w} and X_{a_l} within \mathcal{D}_{pref} .

4 Experiments

4.1 Implementation

We follow (Sun et al., 2024b) to construct our model architecture, including the visual encoder, image projector, prototype extractor, and the instructional LLM. Our proposed self-training pipeline involves two stages. In stage 1, we continually fine-tune the policy model π_θ , initialized from (Sun et al., 2024b) on instructional data with global batch size as 128. During training, we insert LoRA (Hu et al., 2022) with $rank = 128$ and $\alpha = 256$ into the language model (LLM-LoRA) and LoRA with $rank = 32$ and $\alpha = 64$ into the vision encoder (ViT-LoRA). We optimize the model using AdamW (Loshchilov and Hutter, 2019) optimizer for one epoch by setting the learning rate to 2×10^{-4} for LoRA, and 5×10^{-5} for the other

Table 2: Comparison with other methods on three benchmarks. Open questions are evaluated by Recall and F1 score, and closed questions are evaluated by accuracy. All models are using 7B LLM. STLLaVA-Med w/o DPO is the ablated version of our final model. Notably, LLaVA-Med was trained on the original Med-60k-IM (Li et al., 2023), which has 20k more samples than the Med-IM we used in this work due to image unavailability.

Dataset	Method	VQA-RAD			SLAKE			PVQA		
		Recall	F1 Score	Closed	Recall	F1 Score	Closed	Recall	F1 Score	Closed
w/o Med-IM	GPT-4o (OpenAI, 2024)	51.60	9.23	63.97	59.06	8.90	71.63	24.14	3.29	75.97
	LLaVA-v1.5 (Liu et al., 2023a)	23.63	9.53	50.74	35.23	8.84	52.16	11.85	2.73	52.76
	SQ-LLaVA (Sun et al., 2024b)	23.91	6.29	52.57	40.04	9.65	57.45	11.24	2.63	53.73
	Med-Flamingo (Moor et al., 2023)	10.32	10.37	52.21	8.46	7.67	37.02	1.23	1.24	45.59
	PMC-VQA (Zhang et al., 2023)	6.26	5.68	41.54	7.29	6.92	33.89	1.02	1.01	40.10
Med-IM	LLaVA-Med (Li et al., 2023)	32.68	8.65	59.56	40.84	8.21	46.88	12.03	2.47	55.23
	STLLaVA-Med w/o DPO	33.81	10.37	59.16	40.13	10.97	55.53	10.38	2.68	52.05
	STLLaVA-Med	37.12	10.83	60.35	46.69	11.46	57.69	11.92	2.72	52.90

layers. In stage 2, we fine-tune π_θ on the auto-generated preference dataset. Similar to stage 1, we utilize LoRA for light-weight training. We optimize the model using AdamW (Loshchilov and Hutter, 2019) optimizer for one epoch by setting the learning rate to 5×10^{-6} for LoRA, and 2×10^{-5} for the other layers. We train the model on 4 A100s for 10 hours.

4.2 Preference Data Generation

In previous research (Rafailov et al., 2023), preference data were annotated by human annotators. In contrast, this work employs GPT-4o (OpenAI, 2024) as a simulated expert to classify the answers generated by STLLaVA-Med. As shown in Fig. 3, we provide the detailed prompt design for GPT-4o to label the answers with either *win* or *loss*. Due to its multi-modal understanding capabilities, GPT-4o can directly take images as input. In Appendix B.2, we provide qualitative results about the preference data generated by STLLaVA-Med.

4.3 Datasets and Metrics

Datasets. We conduct experiments on the widely-used medical VQA benchmark dataset VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PVQA (He et al., 2020). Specifically, VQA-RAD contains QA pairs generated by clinicians, where the images are evenly distributed over the head, chest, and abdomen. Questions are categorized into 11 categories: abnormality, attribute, modality, organ system, color, counting, etc. SLAKE is a Semantically-Labeled Knowledge-Enhanced dataset for medical VQA. The original dataset contains Chinese and English QA, but we only consider the English subset in our study. Besides, SLAKE includes richer modalities and covers more human body parts than the currently available dataset. PathVQA is a dataset of pathology

images. Each image has questions about multiple aspects, such as location, shape, color, appearance, etc. Overall, the medical questions are categorized into two types: **open-ended** questions such as why, what, how, where, etc., and **closed-ended** questions with one-word answers (yes/no).

Evaluation Metrics. We report accuracy for closed questions. For open-ended questions, we compute recall as the proportion of correctly predicted words out of the reference sentence, and F1 score as a balance metric between recall and precision.

4.4 Overall Performance

The evaluation results in Table 2 are divided into three sections based on rows. We compile all the experiments locally with a single run. The first row indicates the upper bound of zero-shot medical performance; the next two rows and the last three rows reflect the performance of LVLMS trained without and with medical data. As shown in Table 2, even without pre-training on medical data, STLLaVA-Med achieves competitive performance with LLaVA-Med only after fine-tuned instructional data Med-IM. After DPO training, we observed a performance improvement over open-ended questions, demonstrating the effectiveness of supervised preference optimization. Also, in Fig. 4, the answer of STLLaVA-Med has higher relevance, more detail, and accuracy than STLLaVA-Med w/o DPO, proving that self-training on the auto-generated preference dataset controls the model following pre-defined preference. See Appendix B.1 for more results. Moreover, we have several observations based on the results in Table 2.

1) Medical image-text alignment is unnecessary for LVLMS. As we observed, fine-tuning a pre-trained general-purpose LVLMS (e.g., LLaVA-V1.5 (Liu et al., 2023a), SQ-LLaVA (Sun et al., 2024b)) on a small set of medical instruction data

Prompt GPT-4o to score each of the answers

```
messages = [{"role": "User", "content": """}
Text: {
[Question]\nWhat is the condition of the wall of the mass?\n\n
[Answer 1]\nThe wall of the mass appears to be thickened, but it is ... \n\n[End Answer 1]\n\n
[Answer 2]\nThe wall of the mass appears to be hypointense, which mean... \n\n[End Answer 2]\n\n
[System]\n We would like to request your feedback on the quality of two answers in response to the
user question displayed above. The user asks the question on observing an image. For your
reference, you will see an image. \nPlease rate the helpfulness, relevance, accuracy, level of details
of the answers. Each answer receives an overall score on a scale of 1 to 10, where a higher score
indicates better overall performance.\nPlease first output a single line containing only two values
indicating the scores for Answer 1 and 2, respectively. The two scores are separated by a space. }
Image:{<image>}"}]
```

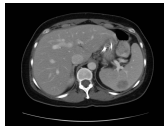
Figure 3: Prompt for GPT-4o to grade the answers generated by STLLaVA-Med from stage 1. The answer with the higher score will be designated as the winning response, while the other will be classified as rejected.

Question: What is abnormal about the spleen?

STLLaVA-Med w/o DPO

The spleen appears to be enlarged and has multiple **hypodense lesions**. These lesions are abnormal growths or fluid-filled sacs within the spleen.

GT: Hypodense lesion



STLLaVA-Med

The spleen appears to be enlarged and has multiple **hypodense lesions**. Hypodense lesions are areas that appear darker on a computed tomography (CT) scan, indicating that they have a lower density compared to the surrounding tissue. This could be due to various reasons, such as infection, inflammation, or other pathological changes.

Figure 4: Qualitative evaluation of methods *w* and *w/o* preference revelation.

achieves the same performance as fully training on medical data (LLaVA-Med). This suggests that general-purpose LVLM with strong vision-language alignment can be easily adapted to medical tasks after light-weight fine-tuning.

2) High-quality medical instruction data can further improve STLLaVA-Med, by enriching the auto-generated data's diversity, complexity, and professionalism.

5 Conclusion

This work has proposed a self-training vision-language assistant for medicine (STLLaVA-Med), a novel training pipeline designed to enhance the data efficiency of training LVLMs for medical tasks. Our approach prompts the policy model to self-generate instruction-answer pairs and label them by a larger language model, such as GPT-4o, for preference optimization. This process aims to enhance the medical reasoning capabilities of a smaller vision-language model, reducing the reliance on extensively annotated medical data and alleviating human experts in the medical field. Experimental results on three benchmarks demonstrate that

STLLaVA-Med achieves exceptional medical reasoning capabilities using medical data at a minimum level. We aspire for our work to inspire future research aimed at enhancing the efficiency of training LVLMs in broad medical domains.

6 Limitations

Although the proposed approach improves medical reasoning ability, the performance of self-training is highly dependent on the quality and relevance of the auto-generated medical instructional data. This indicates that we still need the instructional data from stage 1 training to cover a wider range of medical tasks and professional expertise, which may still be difficult to collect for some diseases or some types of medical images. In addition, GPT4o may become inevitable bias when annotating preference data. To address this, we may use another SOTA LMM, such as Gemini-pro (Team et al., 2023), or include a medical expert in the loop to co-supervise the preference data collection process.

7 Ethics Statements

We conducted experiments and analysis on public datasets, PMC-15M, VQA-RAD, SLAKE, and PVQA, where all medical images and texts were de-identified, ensuring the privacy and confidentiality of patients. While our method reduces the need for extensive labeled datasets, its outputs are still machine-generated, requiring critical human oversight when used in clinical decision-making.

Acknowledgments

This work is in part supported by NIH award R01NR018301.

References

- OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, et al. 2023. Gpt-4 technical report.
- Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. 2023. A general theoretical paradigm to understand learning from human preferences. *AISTAS*.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. In *ArXiv*.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Christopher Olah, Benjamin Mann, and Jared Kaplan. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*.
- Lin Chen, Jinsong Li, Xiao wen Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. In *ArXiv*.
- Yihe Deng, Pan Lu, Fan Yin, Ziniu Hu, Sheng Shen, James Zou, Kai-Wei Chang, and Wei Wang. 2024. Enhancing large vision language models with self-training on image comprehension.
- Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori Hashimoto. 2023. AlpacaFarm: A simulation framework for methods that learn from human feedback. *NeurIPS*.
- Xuehai He, Yichen Zhang, Luntian Mou, Eric P. Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *ArXiv*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.
- Yutao Hu, Tian-Xin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. *ArXiv*.
- Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*.
- Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Fang Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*.
- Haotian Liu, Chunyu Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. In *ArXiv*.
- Haotian Liu, Chunyu Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.
- Viraj Mehta, Vikramjeet Das, Ojash Neopane, Yijia Dai, Ilija Bogunovic, Jeff Schneider, and Willie Neiswanger. 2023. Sample efficient reinforcement learning from human feedback via active exploration. *ArXiv*.
- Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yashodhara Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. 2023. Med-flamingo: a multimodal medical few-shot learner. *ArXiv*.
- OpenAI. 2024. [Gpt-4o system card](#).
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *NeurIPS*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*.

- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *ArXiv*.
- Guohao Sun, Yue Bai, Xueying Yang, Yi Fang, Yun Fu, and Zhiqiang Tao. 2024a. Aligning out-of-distribution web images and caption semantics via evidential learning. In *Proceedings of the ACM Web Conference 2024*.
- Guohao Sun, Can Qin, Jiamian Wang, Zeyuan Chen, Ran Xu, and Zhiqiang Tao. 2024b. Sq-llava: Self-questioning for large vision-language assistant. In *ECCV*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. In *ArXiv*.
- Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuvveer Rao, and Zhiqiang Tao. Text is mass: Modeling as stochastic embedding for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Siyuan Wang, Zheng Liu, and Bo Peng. 2023. A self-training framework for automated medical report generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16443–16449.
- Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, et al. 2024. xgen-mm (blip-3): A family of open large multimodal models. *arXiv preprint arXiv:2408.08872*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhao Chen. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *ArXiv*.
- Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *ArXiv*.
- Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiao wen Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing llms through hallucination-aware direct preference optimization. *ArXiv*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haoteng Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.
- Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. 2024. Aligning modalities in vision large language models via preference fine-tuning. *ArXiv*.

A Related Work

A.1 Large Vision-Language Model

As the field of Large Language Models (LLMs) and instruction tuning undergoes rapid advancements, the academic research community is increasingly focusing on integrating visual information into LLM frameworks to enhance vision-language modeling (Radford et al., 2021; Li et al., 2022; Sun et al., 2024a; Wang et al.; Xue et al., 2024). This emerging research area has seen the development of various methodologies, building on the foundational work of CLIP (Radford et al., 2021) and diverse LLM architectures such as Vicuna (Zheng et al., 2023), Llama2 (Touvron et al., 2023), and Qwen-VL (Bai et al., 2023). Notably, LLaVA (Liu et al., 2023b) pioneered the integration of an LLM with a CLIP vision encoder to create a vision-language model, demonstrating significant capabilities in image-text dialogue tasks through strategies of pre-training alignment and targeted instruction tuning. Subsequent research has focused on refining visual instruction tuning by improving the quality and diversity of datasets used during the pre-training and fine-tuning phases. Building upon these advancements, recent studies, including LLaVA-v1.5 (Liu et al., 2023a) and

Question: What are positively charged, thus allowing the compaction of the negatively charged DNA?

STLLaVA-Med w/o DPO

In the context of the image, **histones are positively charged**, which means they have a positive charge at physiological pH. This positive charge allows them to interact with the negatively charged DNA, helping to compact and organize the DNA into a more compact structure.

GT: the histone subunits



STLLaVA-Med

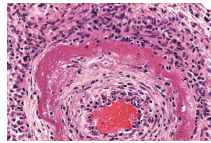
In the context of the image, **histones are positively charged**, which means they have a positive charge at physiological pH. This positive charge allows them to interact with the negatively charged DNA, helping to compact and organize the DNA into a more compact structure. Histones are essential components of chromatin, which is the complex of DNA and histones that makes up the nucleosome, a fundamental unit of chromosome organization.

Question: What does the wall of the artery show with protein deposition and inflammation?

STLLaVA-Med w/o DPO

In the context of this image, the presence of protein deposition and inflammation in the arterial wall may indicate an underlying vascular condition or injury that requires further investigation and appropriate treatment.

GT: a circumferential bright pink area of necrosis



STLLaVA-Med

In the histopathological image, the wall of the artery shows protein deposition and inflammation. Protein deposition refers to the accumulation of proteins in the arterial wall, which can be a sign of injury, inflammation, or other pathological processes. Inflammation is the body's response to injury or infection, and it can cause redness, swelling, and pain in the affected area. In the context of this image, the presence of protein deposition and inflammation in the arterial wall may indicate an underlying vascular condition or injury that requires further investigation and appropriate treatment.

Figure 5: Qualitative evaluation of methods *w* and *w/o* preference revelation.

ShareGPT4V (Chen et al., 2023), have achieved notable success in general vision-language comprehension, showcasing their ability to handle complex question-answering tasks. This progression underscores the importance of sophisticated data handling and model-tuning strategies in developing effective vision-language models.

Table 3: Medical data statistics of training.

Method	#Images	#QA-Pairs
Med-IM	56708	164231
VQA-RAD	313	3064
SLAKE	546	11934
PVQA	37452	26034

Alignment fine-tuning. Following supervised fine-tuning (SFT), alignment fine-tuning has emerged as a key method to further enhance the performance of Large Language Models (LLMs) by aligning them with human preferences (Ouyang et al., 2022). Initial approaches utilized on-policy reinforcement learning (RL) methods, such as proximal policy op-

timization (PPO) (Schulman et al., 2017), to train a reward model based on preference data (Bai et al., 2022). The introduction of direct policy optimization (DPO) (Rafailov et al., 2023; Dubois et al., 2023; Azar et al., 2023; Mehta et al., 2023) has marked a significant shift towards direct learning from human preferences, bypassing the need for an explicit reward model. Another effective strategy is iterative preference fine-tuning, which repeatedly optimizes the model on newly generated preference pairs in successive iterations, thereby improving performance. Despite extensive research on alignment fine-tuning for LLMs, the application of these techniques to Large Vision-Language Models (LVLMs) has been comparatively limited. Early attempts (Zhou et al., 2024; Deng et al., 2024) have focused on constructing preference datasets using human-labeled data or GPT-4 generations, followed by fine-tuning with a DPO loss.

B Experiments

B.1 Additional Results

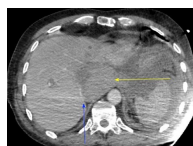
In addition to evaluating zero-shot performance, we conducted experiments involving fine-tuning the model on downstream tasks. To maintain task generalizability and domain specificity, we compiled a new medical instructional dataset, combining Med-IM (Li et al., 2023) with QA pairs from the training sets of VQA-RAD, SLAKE, and PVQA. Table 3 details the number of medical images and QA pairs within each dataset. After fine-tuning the models on this visual instruction dataset, we observed a clear improvement in downstream task performance, as shown in Table 4. The performance gap between the baseline model LLaVA-v1.5 and the proposed STLLaVA-Med demonstrates the effectiveness of the self-training pipeline.

Additionally, the improvement between STLLaVA-Med without DPO and STLLaVA-Med illustrates the effectiveness of preference alignment within the self-training pipeline. However, we found this improvement is not as significant as the improvement over zero-shot scenario. One explanation is the inconsistency between our designed preference and the ground truth preference. For VQA-RAD, SLAKE, and PVQA, the ground truth are short phrases, but the preference we are trying to optimize is *detailed* and *relevance*. This gives us an insight that the human expert should be involved in future medical tasks evaluation.

Table 4: Comparison of fine-tuning performance on three benchmarks. Open questions are evaluated by Recall and F1 score, and closed questions are evaluated by accuracy. All models are using 7B LLM. STLLaVA-Med w/o DPO is the ablated version of our final model.

Dataset	Method	VQA-RAD			SLAKE			PVQA		
		Recall	F1 Score	Closed	Recall	F1 Score	Closed	Recall	F1 Score	Closed
Med-IM+	LLaVA-v1.5 (Liu et al., 2023a)	43.44	36.41	70.59	52.76	46.98	64.18	35.91	35.47	91.15
VQA-RAD+	STLLaVA-Med w/o DPO	52.07	45.38	75.74	56.10	50.77	67.31	38.05	37.76	92.13
SLAKE+PVQA	STLLaVA-Med	52.60	45.92	76.10	57.37	50.84	67.31	38.30	38.00	92.13

Question: Is there any infiltration of the lesser curvature of the stomach or duodenum?



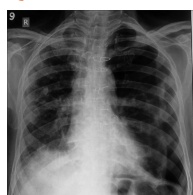
Answer 1: According to the axial abdominal CT scan, there is no infiltration of the lesser curvature of the stomach or duodenum.

 Answer 1

Answer 2: No, the CT scan does not show any infiltration of the lesser curvature of the stomach or duodenum.

Y
Answer 2

Question: What could be the possible cause of bilateral infiltrates?



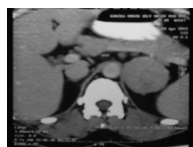
Answer 1: Bilateral infiltrates on a chest X-ray can be caused by various conditions, including infections (such as pneumonia), inflammation, or other lung diseases. Some possible causes are bacterial, viral, or fungal infections, aspiration of foreign material, or autoimmune disorders. It is important to consider the patient's clinical history, symptoms, and additional diagnostic tests to determine the specific cause of the bilateral infiltrates and provide appropriate treatment.

 Answer 1

Answer 2: Bilateral infiltrates on a chest X-ray can be caused by various factors, such as infections (like pneumonia), inflammation, heart failure, or lung diseases.

Y
Answer 2

Question: What kind of lesion is it?



Answer 1: The lesion appears to be a multilocular, cystic retroperitoneal tumor. It is a fluid-filled mass that is situated in the retroperitoneal space, which is the area behind the peritoneum, the membrane that lines the abdominal cavity. The lesion is divided into multiple septations, which are thin walls or partitions within the cystic structure.

 Answer 1

Answer 2: The lesion appears to be a soft tissue heterogeneously enhancing lesion with a nodular appearance. It could be suggestive of a reactive lymph node or another type of lesion.

Y
Answer 2

Figure 6: Preference data visualization. The win and loss answer were classified by GPT-4o.

B.2 Qualitative Results

In Table 2, we have observed a clear improvement of model performance after preference optimization. In Fig. 5, we provide more qualitative results of medical VQA. As can be seen, STLLaVA-Med follows human preference by generating more detailed and accurate answers than the model without DPO fine-tuning. Fig. 6 provides example samples of the preference data generated by STLLaVA-Med and GPT-4o. From these three samples, we find that the chosen answers contain more detail and in-depth analysis, aligning with human preference.