Biomedical Visual Instruction Tuning
with Clinician Preference Alignment

Hejie Cui[1]*, **Lingjun Mao**[3]*, **Xin Liang**[3], **Jieyu Zhang**[4],
Hui Ren[5,6], **Quanzheng Li**[5,6], **Xiang Li**[5,6]†, **Carl Yang**[2]†

[1] Stanford University [2] Emory University [3] University of California, Berkeley
[4] University of Washington [5] Massachusetts General Hospital [6] Harvard Medical School

Abstract

Recent advancements in multimodal foundation models have showcased impressive capabilities in understanding and reasoning with visual and textual information. Adapting these foundation models trained for general usage to specialized domains like biomedicine requires large-scale domain-specific instruction datasets. While existing works have explored curating such datasets automatically, the resultant datasets are not explicitly aligned with domain expertise. In this work, we propose a data-centric framework, **Biomed**ical **V**isual **I**nstruction **T**uning with Clinician Preference **Al**ignment (BioMed-VITAL), that incorporates clinician preferences into both stages of generating and selecting instruction data for tuning biomedical multimodal foundation models. First, during the generation stage, we prompt the GPT-4V generator with a diverse set of clinician-selected demonstrations for preference-aligned data candidate generation. Then, during the selection phase, we train a separate selection model, which explicitly distills clinician and policy-guided model preferences into a rating function to select high-quality data for medical instruction tuning. Results show that the model tuned with the instruction data from our method demonstrates a significant improvement in open visual chat (18.5% relatively) and medical VQA (win rate up to 81.73%). Our instruction-following data, models, and code are available at `https://BioMed-VITAL.github.io`.

1 Introduction

Recent advancements in large pre-trained multimodal models, such as GPT-4V [1], have demonstrated impressive performance on various language and vision tasks. However, when directly applied to specialized domains like biomedicine, these models may fall short due to their primary focus on general usage rather than domain-specific expertise [44, 32]. To bridge this gap and adapt general domain models to specialized domains, researchers have explored various techniques. Instruction tuning has emerged as a promising approach, involving the fine-tuning of large foundation models to follow explicit, natural language instructions [43, 29, 48]. These instructions are composed of task-specific prompts and their corresponding response, enabling the models to learn and generalize to a wide range of tasks within the target domain.

Although instruction tuning has proven to be an effective method for adapting models to target domains and performing various downstream tasks, its success heavily depends on large-scale instruction-following datasets. Curating large-scale instructional datasets in specialized domains, such as biomedicine, can be expensive and time-consuming, often requiring significant domain expertise. Previous work proposes to use strong language models to generate instruction data

---

*These authors contributed equally to this work.

†co-corresponding: xli60@mgh.harvard.edu, j.carlyang@emory.edu

automatically, which effectively reduces the need for extensive manual annotation [36]. Such paradigms have successfully been adopted to adapt general domain models to biomedicine. For example, LLaVA-Med [19] developed a framework to instruction-tune biomedical language-vision models with GPT-4 generated instruction-following data. This approach has achieved impressive performance on open-ended visual chat and visual question answering benchmarks, highlighting the potential of using model-generation data in the biomedical domain.

However, existing methods for automatically curating datasets do not explicitly incorporate clinician preferences, which may result in models producing irrelevant or impractical output, limiting their utility in real-world applications [11]. Yet, aligning domain expertise with the process of instruction-following datasets curation is challenging. First, advanced data generators, such as GPT-4V, are often proprietary and not publicly available for alignment tuning. Second, clinician-annotated preference data in the biomedical domain is limited, further restricting effective preference learning. The combination of model opacity and data scarcity creates a significant bottleneck in developing high-quality, expert-aligned instruction-following data for instruction-tuning. This hinders the development of domain-specific models that can effectively incorporate expert preferences and requirements, ultimately limiting their practical utility and real-world impact.

To tackle this challenge, we propose an effective data-centric approach, BioMed-VITAL, that incorporates clinician preference into the process of automatically curating instruction-following data for biomedical visual instruction tuning. As shown in Figure 1, BioMed-VITAL consists of three stages: (1) data generation with demonstrations, (2) data selection with a preference distilled model, and (3) visual instruction-tuning. In data generation, we strategically sample a diverse set of instructions to collect clinician preferences, which are used as demonstrations for GPT-4V-based instructional data generation, guiding the data generation toward producing more clinically relevant and useful instruction-following examples. In the data selection stage, we train a data selection model that distills a mixture of preferences from clinician-annotated and model-annotated data guided by clinician-curated criteria. This model is then used to rank the generated data samples, and the top-ranked samples are selected for visual instruction-tuning.

The contributions of this work are summarized as follows:

- We introduce a data-centric framework BioMed-VITAL, which generates and selects instruction-following data aligned with clinician preference for visual instruction tuning. Evaluation indicates an improved data quality and our instruction-tuned models remarkably improve in both open visual chat (18.5% relatively) and three biomedical VQA benchmarks (win rate up to 81.73%).
- We propose a paradigm involving clinician preference during generation and an effective data selection model based on a mixture of preferences. It is shown that our distilled data selection model excels in matching human preferences compared with judgments of GPT-4.
- To facilitate further study, we release $80k$ clinician preference-aligned instruction-following datasets generated and selected from ours, along with the models instruction-tuned based on them. All resources are publicly available on the website `https://BioMed-VITAL.github.io`.

2   Background

Instruction-Tuning. Instruction tuning has become an effective method for adapting pre-trained language models to a wide range of natural language tasks [50, 42, 41, 45, 12, 31, 35, 6] by providing task-specific instructions and examples. This approach has been further explored in studies like FLAN-T5 [8], LLaMA [37], and LLaMA2 [38], which enables models to understand and follow task-specific instructions without extensive task-specific fine-tuning. Recently, using strong language models to generate instruction data automatically has been proposed to train a high-quality instruction-following model under an academic budget [33, 36, 25]. For example, Stanford Alpaca [36] instruction-tuned LLaMA using `text-davinci-003`-generated instruction-following datasets and achieved competitive performance on various NLP tasks.

Vision-Language Foundation Models in Biomedical Domain. General vision-language foundation models have achieved remarkable success across various domains. Researchers in biomedicine have been actively exploring the adaptation of vision-language foundation models to tackle domain-specific tasks [30, 2, 15, 34]. However, effectively adapting vision-language foundation models to specialized domains such as the biomedical presents challenges, particularly due to limited training data. To overcome this challenge, our work aims to establish a data-centric method that aligns domain
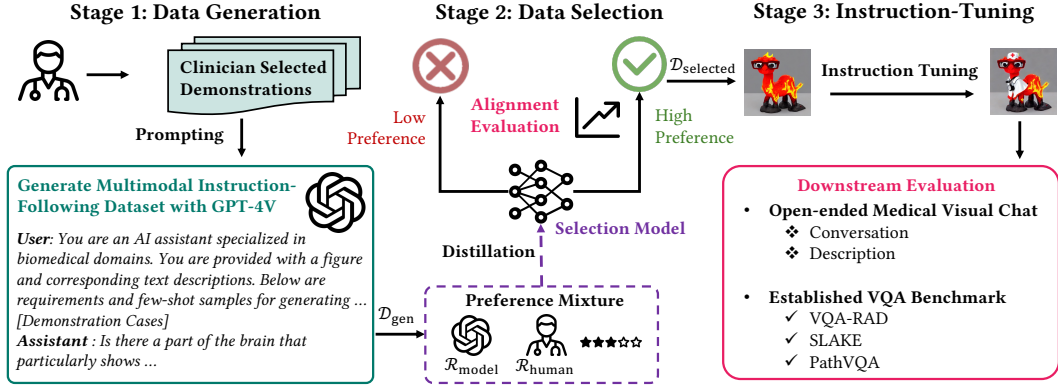
Stage 1: Data Generation

Clinician Selected Demonstrations

Prompting

Generate Multimodal Instruction-Following Dataset with GPT-4V

User: You are an AI assistant specialized in biomedical domains. You are provided with a figure and corresponding text descriptions. Below are requirements and few-shot samples for generating ...
[Demonstration Cases]
Assistant : Is there a part of the brain that particularly shows ...

$\mathcal{D}_{\text{gen}}$

Stage 2: Data Selection

Low Preference

Alignment Evaluation

High Preference

Selection Model

Distillation

Preference Mixture

$\mathcal{R}_{\text{model}}$ $\mathcal{R}_{\text{human}}$ ★★★☆☆

$\mathcal{D}_{\text{selected}}$

Stage 3: Instruction-Tuning

Instruction Tuning

Downstream Evaluation

- **Open-ended Medical Visual Chat**
  - ❖ Conversation
  - ❖ Description
- **Established VQA Benchmark**
  - ✓ VQA-RAD
  - ✓ SLAKE
  - ✓ PathVQA

Figure 1: Overview of **Biomed**ical **V**isual **I**nstruction **T**uning with Clinician Preference **Al**ignment (BioMed-VITAL). Clinician preferences are infused in the 1. data generation and 2. selection stages.

expertise from clinicians with the instructional data for instruction-tuning, which generates and selects instruction-following datasets that are aligned with clinician preference.

3 Clinician-Aligned Biomedical Visual Instruction Tuning

Figure 1 presents an overview of the proposed framework BioMed-VITAL, consisting of three stages: (1) data generation with diverse expert-selected demonstration, (2) data selection with a distilled selection model trained with mixed preferences, and (3) instruction tuning to adapt a general multimodal model for biomedical tasks. The output from the framework includes a clinician preference-aligned instruction-following dataset $\mathcal{D} = \{(I_i, C_i, \mathbf{Q}_i, \mathbf{A}_i)\}_{i=1}^N$ and instruction-tuned models based on it. I_i represents the i-th biomedical image; C_i is the caption and inline-mentions associated with the i-th image; $\mathbf{Q}_i = \{\mathcal{Q}_{ij}\}_{j=1}^{n_i}$ contains n_i instructions, where j represents the j-th instruction for the i-th image-text sample; $\mathbf{A}_i = \{\mathcal{A}_{ij}\}_{j=1}^{n_i}$ contains n_i responses, each corresponding to $\mathcal{Q}_{ij}$; and N is the total number of samples in the dataset.

3.1 Stage 1: Data Generation with Diverse Expert-Selected Demonstration

Large pre-trained models have shown strong in-context learning capabilities by learning from a few presented examples and mimicking when generating responses. In BioMed-VITAL, we use the GPT-4V model as the generator. To incorporate clinician preference into the data generation process, we first select a diverse set of samples for clinicians to annotate. Clinician-selected QA pairs are used as few-shot demonstrations for GPT-4V to generate instruction-following data at scale.

Diverse few-shot demonstration selection. We employ a strategic sampling approach to ensure the diversity and representatives of the demonstrations for the generator. For each sample (I_i, C_i) in the dataset, the image and text representations are extracted using BiomedCLIP [47], then we perform K-means clustering on these representations to cluster the samples into K distinct categories, denoted as $\mathcal{D}_1, \mathcal{D}_2, ..., \mathcal{D}_K$. From these clusters, we uniformly select a subset $S = (I_i, C_i)_{i=1}^M$ with total M samples that have relatively complex captions and inline mentions. For each selected sample $(I_i, C_i) \in S$, we use GPT-4V to generate conversations that present instructions $\mathcal{Q}_i$ and two candidate responses A_i^1, A_i^2 for each instruction [3]. During the human annotation, each clinician is presented with a carefully selected set of M questions with two response candidates and asked to choose the preferred one A_{ij}^{pref} between the two, select both if two responses are equally good, or deselect both to drop this instruction [4]. Three clinicians participated in our annotation. The Fleiss' kappa on the three clinician annotations is 0.736, indicating a good agreement among the three annotators. For these cases where disagreement happened, we applied majority voting to make the final decision. The resulting annotation $\mathcal{R}_{\text{human}}$ contains the selected preferences from clinicians.

---

[3]The prompt can be found in Appendix B Figure 6.

[4]See Appendix B Figure 7 for the annotation protocol.

Instruction-following data generation with GPT-4V. Using the clinician-selected data, we employ GPT-4V as the generator to simulate the instructional dataset. During each API call, we randomly select 2 samples for each of the 5 modalities from $\mathcal{D}_{\text{pref}}$ as few-shot demonstrations and append them to the language prompts. The full prompt can be referred to in Appendix A Figure 5. Compared with previous methods, our generated dataset $\mathcal{D}_{\text{gen}} = \{(I_i, C_i, \mathbf{Q}_i, \mathbf{A}_i)\}_{i=1}^{N}$ incorporate visual input and is further guided with selected clinician demonstrations.

3.2 Stage 2: Distilling Mixed Clinician Preference for Data Selection

While $\mathcal{D}_{\text{gen}}$ is directly usable to instruction-tune, it may still include samples that can introduce noise or bias or are irrelevant to the real needs of clinicians. In the second stage of BioMed-VITAL, we train a data selection model that learns to select instruction data aligned with expert preference.

Preference data from two resources. Collecting human preference data from domain experts such as clinicians is expensive and time-consuming. Thus, the available annotation data is usually on a small scale. A recent paradigm involves using LLMs as judges, which have been shown to match human preferences effectively [51]. We consider a data mixing schema to distill preference into a local model for data selection. Our preference data comes from two resources, from humans and from models: (1) human preference from the preference annotation $\mathcal{R}_{\text{human}}$ in stage 1, where each question $\mathcal{Q}_{ij}$ is paired with two candidate answers A_{ij}^1, A_{ij}^2, with A_{ij}^{pref} annotated as the preferred one. (2) model-based preference: to generate reliable model-based ratings, we first collect a set of clinician-curated factors for data quality evaluation, such as missing information, recognition errors, lack of medical precision, insufficient depth, valueless questions, etc. With these clinician-curated criteria, we use GPT-4V as a judge to score a randomly sampled set of data from 0 to 10. The detailed prompt can be referred to in Appendix B Figure 8. The resulting self-evaluated ratings, $\mathcal{R}_{\text{model}}$, provide additional preference data and address the scalability issue related to human annotation.

Distill clinician preference to a selection model. Next, we train a data selection model with the preference data, which is designed to identify and remove low-quality samples from the generated dataset and preserve only the most accurate and clinically relevant examples for instruction tuning. We use BiomedCLIP [47] as the backbone, followed by an MLP head to perform binary prediction tasks on good/bad ratings of data samples. Pairwise ranking loss is used as the training objective: given a pair of candidate samples x_i and x_j, along with their corresponding annotated preferences $\mathcal{R}_i$ and $\mathcal{R}_j$, the objective is formulated as a pairwise classification:

$$\mathcal{L}_Q = -z_i \log \sigma \left(f(x_i) \right) - z_j \log \sigma \left(f(x_j) \right), \tag{1}$$

where σ represents the sigmoid function, and $f(\cdot)$ denotes the rating function learned by the model. The values of z_i and z_j are determined by comparing the preference annotation:

$$(z_i, z_j) = \left\{ \begin{array}{ll} (1, 0), & \mathcal{R}_i \geq \mathcal{R}_j \\ (0, 1), & \mathcal{R}_i < \mathcal{R}_j \end{array} \right. . \tag{2}$$

By minimizing the pairwise classification loss, the data selection model learns to predict the likelihood of a sample being labeled as 1 within a sampled pair, by assigning higher scores to samples with higher preference and lower scores to samples with lower preference.

Preference mixing strategy during training. We mix two sources of preference data in each batch during training. In Eq (1), each x_i and x_j can be either human-annotated preferences from $\mathcal{R}_{\text{human}}$, or two samples with model-based ratings $\mathcal{R}_i$ and $\mathcal{R}_j$ from $\mathcal{R}_{\text{model}}$. To address the scalability difference between the two resources, we introduce an adaptive contribution mechanism by incorporating an adjustable sample weight $w_{i,j}$ into Eq (1):

$$\mathcal{L}_Q = -w_{i,j} \left(z_i \log \sigma \left(f(x_i) \right) + z_j \log \sigma \left(f(x_j) \right) \right), \tag{3}$$

where $w_{i,j}$ allows for an adjustable preference contribution from human or model during training.

Data selection with distilled selection model. We apply the trained data selection model to the generated dataset $\mathcal{D}_{\text{gen}}$ and observe F1@K and Precision@K curves to determine the threshold for data selection. To balance data quality and diversity, we first cluster all the data samples into K groups and uniformly select top-ranked data in each group to compose the final instruction-following dataset, denoted as $\mathcal{D}_{\text{selected}}$, which contains the most informative, accurate, and clinically relevant examples. More empirical decisions during selection are discussed in Section 4.2.

3.3 Stage 3: Instruction-Tuning

Following LLaVA-Med [19], we continue training the LLaVA [26, 24] model on our curated instruction-following dataset $\mathcal{D}_{\text{selected}}$. The instruction tuning objective for model θ is to minimize the negative log-likelihood of the target $\mathbf{A}_i$ given input image I_i, caption C_i, question $\mathbf{Q}_i$,

$$\mathcal{L}_{IT} = - \sum_{i=1}^{|\mathcal{D}_{\text{selected}}|} \log p(\mathbf{A}_i | I_i, C_i, \mathbf{Q}_i, \theta). \tag{4}$$

4 Experiments

4.1 Dataset and Experiment Details of BioMed-VITAL

We follow the setup of Li et al. [47] and utilize image-text pairs from the PMC-15M dataset [47] to generate multi-round QA instructional data. For the data generator, we utilize `gpt-4-vision-preview` API on Azure OpenAI. For the diverse few-shot demonstration selection, we set K to 60 and M to 300 for simplicity. For the data selector, we use BiomedCLIP [47], which is trained for 6 epochs with a learning rate of 1e-4. For instruction-tuning, we use `llava-v1.5-13b` as the backbone. Following the LLaVA-Med approach [19], the model is first trained with biomedical concept alignment; subsequently, it is instruction-tuned using the selected dataset from the second stage, utilizing a multi-turn dialogue setup [25]. The instruction-tuning process is carried out for 3 epochs with a learning rate of 2e-5, trained and tested with 2 NVIDIA A100 GPUs.

4.2 Alignment Evaluation of the Data Selection Model

Table 1: Varying preference mixture strategy.

Mixture Strategy	Rank-based Metrics (%)			
	ACC ↑	AUC ↑	MR ↓	MAP ↑
only $\mathcal{R}_{\text{human}}$	55.89	55.99	46.91	56.21
only $\mathcal{R}_{\text{model}}$	54.76	54.64	47.67	55.25
mix, $w_{\mathcal{R}_{\text{human}}}/w_{\mathcal{R}_{\text{model}}} = 1$	61.61	61.90	44.04	62.22
mix, $w_{\mathcal{R}_{\text{human}}}/w_{\mathcal{R}_{\text{model}}} = 5$	58.63	58.22	45.87	62.29
mix, $w_{\mathcal{R}_{\text{human}}}/w_{\mathcal{R}_{\text{model}}} = 10$	59.38	59.14	45.39	59.20
mix, $w_{\mathcal{R}_{\text{human}}}/w_{\mathcal{R}_{\text{model}}} = 50$	59.67	59.80	45.09	63.27
mix, $w_{\mathcal{R}_{\text{human}}}/w_{\mathcal{R}_{\text{model}}} = 100$	62.05	62.30	43.84	61.63
mix, $w_{\mathcal{R}_{\text{human}}}/w_{\mathcal{R}_{\text{model}}} = 200$	60.91	61.23	44.37	59.55
mix, $w_{\mathcal{R}_{\text{human}}}/w_{\mathcal{R}_{\text{model}}} = 300$	63.64	63.12	43.43	63.00
mix, $w_{\mathcal{R}_{\text{human}}}/w_{\mathcal{R}_{\text{model}}} = \mathbf{400}$	**66.72**	**66.32**	**41.83**	**64.47**
mix, $w_{\mathcal{R}_{\text{human}}}/w_{\mathcal{R}_{\text{model}}} = 500$	62.85	63.06	43.46	65.00
mix, $w_{\mathcal{R}_{\text{human}}}/w_{\mathcal{R}_{\text{model}}} = 600$	56.30	56.07	46.95	60.25

Preference mixture. In our experiment, we train the selection model with a proportion of 1:400 of human and model preferences to reflect the scalability gap. To find the optimal balance between the two resources, we adjust the ratio of sample weight for human-annotated and model-generated preference $w_{\mathcal{R}_{\text{human}}}/w_{\mathcal{R}_{\text{model}}}$ when applying Eq (3) to train the selection model and observe the performance with varying mixture strategies on a clinician-annotated test set. We also conducted experiments to train the data selection model using only human-generated preference $\mathcal{R}_{\text{human}}$ or only model-generated ratings $\mathcal{R}_{\text{model}}$. The results are summarized in Table 1.

It shows that when trained with only $\mathcal{R}_{\text{human}}$ or $\mathcal{R}_{\text{model}}$, the selection model's performance is inferior to the stratified mixture of both preference data. This indicates that while high-quality, the limited amount of annotations is insufficient for robust data selection model training. In comparison, stratified mixing of $\mathcal{R}_{\text{human}}$ and $\mathcal{R}_{\text{model}}$ significantly improves performance compared to using only $\mathcal{R}_{\text{model}}$. The best performance is achieved with a ratio of 400, which happens to balance the contribution of human and model-annotated preference in the total training loss. These findings strongly support our approach of preference mixing strategy, which effectively balances the high-quality but limited clinician annotations with scalable model-based annotations, resulting in a more accurate selection model that successfully distillates clinician preferences while minimizing human effort in annotation.

Alignment with human preference versus GPT-4. To compare the preference evaluation ability of our trained data selection model versus the GPT-4 model, we calculate the correlation between the ratings generated from both models with gold clinician-annotated preference. The results in Figure 2 (left panel) indicate a better alignment of our trained selection model over the GPT-4 model.

Selecting top K ranked samples. We observe the F1 and Precision performance curves on the ranking list from the score model by varying the top K percentiles to determine the optimal proportion of top-ranked data to select. As illustrated by Figure 2 (right panel), we identify three critical percentiles: 10%, 50%, and 80%, where the performance either reaches a local peak or plateaus afterward, indicating that further incorporating data on the ranking list would not yield significant improvements. The top 50% and the top 80% are selected because they demonstrate similar F1 scores and precision
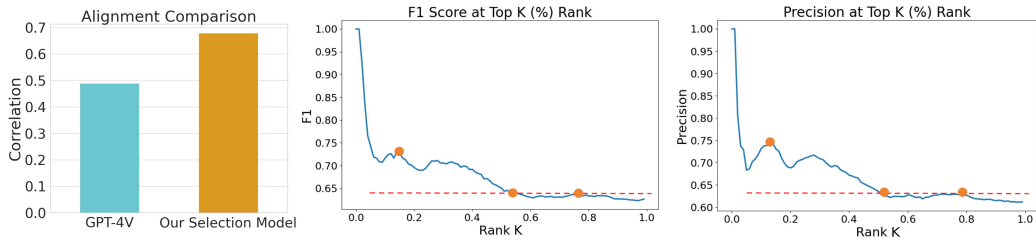
Figure 2: Left: Comparison of human preference alignment between GPT-4V and our selection model. Right: F1 and precision for varying top K percentile samples ranked by the selection model.

when compared to human annotations. This allows us to understand the influence of the scaling law in model training. Additionally, we include the top 10% threshold as it represents a subset of data that strikes a balance between quality and quantity to help better understand the effectiveness of data selection. Consequently, we select datasets corresponding to these critical percentiles for visual instruction tuning, ensuring that the models learn from high-quality, clinician-preferred samples.

4.3 Downstream Evaluation 1: Open-Ended Medical Visual Chat

To evaluate the model's ability to engage in dialogue-like interactions and provide coherent responses, we evaluate the model with open-ended visual chat, where the trained language models are prompted to respond to given questions based on the provided images and texts in a multi-round manner.

Dataset and evaluation paradigm. For the evaluation dataset, we use 50 unseen image and caption pairs with 193 question-answer pairs collected by the LLaVA-Med [19] authors. The questions are divided into two types: (1) Conversation questions, which require the model to engage in dialogue-like interaction, understand the context and provide relevant responses. For example, given an image of a chest X-ray, a conversation question might ask, "What abnormalities do you see in this X-ray image?" (2) Description questions, which focus on detailed descriptions or explanations based on visual and textual input. For instance, a description question for a histology image could be, "Describe the morphological features of the cells in this histology slide."

Open-ended visual chat can be challenging to evaluate, with traditional NLP metrics insufficient to capture the semantic and higher-order abstract text aspects. Recent studies have explored using LLMs as evaluators, demonstrating their greater resilience compared to metric-based approaches [3, 7, 27]. Following these, we use GPT-4V as the evaluator. A reference prediction is first generated based on the input context and the given question, which is then provided to assess the responses from various trained models by assigning a relative score on a scale from 1 to 10. A higher score indicates that the model's response is more accurate, relevant, and coherent with respect to the reference prediction.

Table 2: Performance comparison of the instruction-tuned models on open-ended biomedical visual chat. The number followed by "#: " represents the number of testing samples in this category. In the following experiments, N is the number of QA pairs of 60k images.

Model	Data Size	Question Types		Domains					Overall
		Conversation (#:143)	Description (#: 50)	CXR (#: 37)	MRI (#: 38)	Histology (#: 44)	Gross (#: 34)	CT (#: 40)	(#: 193)
LLaVA-Med	N	58.53	56.16	43.97	51.19	60.01	86.49	50.63	57.92
BioMed-VITAL	Top 10% $*N$	64.11	60.05	56.35	52.57	59.02	87.60	62.82	63.06
BioMed-VITAL	Top 50% $*N$	65.95	64.26	55.75	55.57	60.96	94.06	64.70	65.51
BioMed-VITAL	Top 80% $*N$	68.50	**67.65**	55.24	**58.73**	62.65	**101.88**	**67.05**	68.28
BioMed-VITAL	N	**69.73**	65.51	**59.22**	57.39	**67.15**	99.26	63.63	**68.63**
Model Ablation									
BioMed-VITAL[A0]	N	65.38	60.63	63.48	53.82	57.32	92.30	58.16	64.15
BioMed-VITAL[A1]	N	67.82	59.48	59.68	53.98	60.34	97.89	60.74	65.66
BioMed-VITAL[A2]	N	67.53	62.78	60.64	54.62	61.07	98.27	61.21	66.30

Model variants. In addition to comparing our model with the LLaVA-Med baseline, we further investigate the influence of the selected data size on instruction tuning performance and conduct a model ablation study. ♦ To study the impact of data size, we instruction-tune three additional models using datasets selected from the ranking list at three critical percentiles: 10%, 50%, and 80%, as described in Section 4.2 and illustrated in Figure 2. ♦ For the model ablation study, we include three

6

Table 3: Statistics of the benchmark datasets for downstream evaluation on biomedical VQA.

Dataset	VQA-RAD		SLAKE			PathVQA		
	Train	Test	Train	Val	Test	Train	Val	Test
# Images	313	203	450	96	96	2,599	858	858
# QA Pairs	1,797	451	4,919	1,053	1,061	19,755	6,279	6,761
# Open	770	179	2,976	631	645	9,949	3,144	3,370
# Closed	1,027	272	1,943	422	416	9,806	3,135	3,391

variants based on the full BioMed-VITAL model: BioMed-VITAL[A0], which does not incorporate clinician preference alignment in either stage; BioMed-VITAL[A1], which only includes the first stage of clinician-selected demonstrations; and BioMed-VITAL[A1], which only incorporates the second stage of preference distillation. The results of these investigations are summarized in Table 2. Note the numbers are relative scores calculated following LLaVA-Med [19]. When comparing a response generated by a model to a reference answer from GPT-4V, the evaluator is asked to provide scores for both, and a relative score is calculated by dividing the model's score by the reference's score.

Result discussion. For the three-dimensional comparison:

- Baseline comparison: BioMed-VITAL and all its variants consistently outperform the compared method. Even with only the top 10% of selected data, the BioMed-VITAL model surpasses the baseline model trained on the full dataset of size N in both question types, highlighting the effectiveness of our data-centric framework.
- Data size study: When varying the top-ranked percentiles in the data selection process, increasing the dataset size generally improves model performance. Notably, our models trained with fewer data (i.e., 50% and 80% of the dataset) outperform the BioMed-VITAL[A0] and BioMed-VITAL[A1] models, which are trained on the full data size N without data selection. This finding suggests that the second-stage data selection leads to more efficient and effective model tuning, as it focuses on the most informative and relevant examples.
- Model ablation study: Comparing the three model ablations with the full BioMed-VITAL model, we observe that incorporating clinician preference infusion in both the data generation and selection stages leads to improved performance compared to the base model. The full BioMed-VITAL model achieves the best performance, revealing the effectiveness of combining both alignment stages to achieve optimal results. This finding underscores the importance of considering clinician preferences throughout the entire data-centric framework for biomedical visual instruction tuning.

4.4 Downstream Evaluation 2: Performance on Established VQA Benchmarks

Dataset details. We train and evaluate BioMed-VITAL on three widely used biomedical visual question answering benchmarks [19, 39, 49]. The statistics of the datasets are shown in Table 3.

- VQA-RAD [17] is a dataset containing 3,515 question-answer pairs created by medical professionals, along with 315 radiology images. Each image is linked to several questions, which are categorized into 11 types, including abnormality, attribute, modality, organ system, color, counting, object/condition presence, size, plane, positional reasoning, and others. The dataset features a balanced mix of closed-ended (yes/no) and open-ended (one-word or short phrase) answers.
- SLAKE [23] is a comprehensive medical visual question-answering dataset with knowledge-enhancement features. It contains radiology images and diverse question-answer pairs annotated by experienced physicians. SLAKE covers a wide range of modalities and human body parts, such as the brain, neck, chest, abdomen, and pelvic cavity.
- PathVQA [13] focuses on pathology images. Each image is associated with multiple questions that cover various aspects, such as location, shape, color, and appearance. The questions in PathVQA include open-ended questions (e.g., why, what, how, where) and closed-ended questions.

Experimental details. For each benchmark, the model is fine-tuned for 15 epochs with a learning rate of 2e-5. To account for the open-ended nature and expressive diversity of language generation, we report both metrics-based performance and an additional model-based win rate performance. The win rate performance provides a complementary perspective on the model's ability to generate accurate and relevant responses compared to the baseline.

Metric performance. To evaluate the performance metrics, we follow the practice of Li et al. [19] and use accuracy for closed-set questions and recall (the ratio of ground-truth tokens appearing in the generated response) for open-set questions. Table 4 summarizes the metric performance of

Table 4: Metric performance of BioMed-VITAL and compared methods on three VQA benchmarks. Models based on LLaVA are trained with 7b/13b backbone and training sample size of 60K/150K. The largest set 150K combines 10K and 60K provided by LLaVA-Med, plus our curated 80K samples.

Model	VQA-RAD			SLAKE			PathVQA		
	Ref	Open	Closed	Ref	Open	Closed	Ref	Open	Closed
Supervised fine-tuning results from models based on LLaVA (model size, training sample size)									
LLaVA (7B)		50.00	65.07		78.18	63.22		7.74	63.20
LLaVA-Med (7B, 60k)		61.52	84.19		83.08	85.34		37.95	91.21
LLaVA-Med (13B, 60k)		64.58	77.94		84.97	85.58		38.82	92.39
BioMed-VITAL (7B, 60k)		63.46	84.71		85.41	87.26		38.96	92.39
BioMed-VITAL (13B, 60k)		64.88	84.55		87.82	86.54		39.71	91.41
BioMed-VITAL (13B, 150k)		**69.72**	**84.86**		**91.69**	**90.70**		**39.89**	**92.42**
Literature-reported results from representative SoTA methods									
MMQ [9]	53.70		75.80				13.40		84.00
Prefix T. Medical LM [40]				84.30		82.01	40.00		87.00
PubMedCLIP [10]	60.10		80.00	78.40		82.50			
BiomedCLIP [46]	67.60		79.80	82.05		89.70			
M2I2 [21]	66.50		83.50	74.70		91.10	36.30		88.00
MUMC [20]	71.50		84.20	81.50		81.50	39.00		65.10
M3AE [5]	67.23		83.46	80.31		87.82			
CoQAH [16]	30.20		67.50	42.50		73.90			
PMC-CLIP [22]	67.00		84.00	81.90		88.00			

BioMed-VITAL compared to models based on LLaVA, as well as literature-reported results from representative state-of-the-art (SoTA) methods for reference[5]. Among the supervised fine-tuning models based on LLaVA, BioMed-VITAL consistently outperforms the other two, particularly on open-type questions, with the 150k trained model achieving the best. When comparing ours to those reported in the literature from previous methods, it is important to note that some prior methods formulate the problems as classification tasks among answer candidates in the training set, which does not meet the real-world need for open-ended QA. Additionally, some studies report metrics on the open set using different calculations, leading to inconsistencies in comparison. We follow the practice of Li et al. [19] and present the numbers from prior work only as a reference for the open set while including metrics on the closed set for comparison. The results demonstrate that BioMed-VITAL achieves leading performance in most cases, even when compared to methods that employ classification set up for QA despite BioMed-VITAL being in an open, generative manner.

Varying vision-language model backbone and model sizes. We conducted additional experiments with recent SoTA language-vision models, including LLaVA-OneVision [18] and InternVL-1.5 [4], trained on our generated instruction data, and evaluated on three VQA benchmarks. The results are summarized below in Appendix C Table 5. The key findings include: (1) Fine-tuning with our instruct data significantly improves model performance across all three benchmarks, demonstrating the effectiveness of our framework in generating helpful training data. (2) Our approach shows consistent improvement in both open and closed categories compared to fine-tuning with LLaVA-Med datasets, highlighting the benefits of clinician alignment. (3) Performance gains are observed across different model architectures, indicating the generalizability and robustness of our approach.

We also conducted experiments using models of various sizes (7B and 13B parameters). As shown in Appendix C Table 6, our framework consistently outperforms both LLaVA and LLaVA-Med across tasks and model sizes, particularly notable in the open-ended questions for all datasets. This underscores our framework's generalizability across model sizes.

Win rate performance. Recent studies in visual question-answering have highlighted the limitations of token-matching metric evaluation for open-ended language generation tasks and have proposed leveraging model-based win rate evaluation instead [28, 14]. In line with these insights, we adopted a reference-guided win rate evaluation, where GPT-4V is employed as an impartial judge to assess the quality of the responses provided by two compared models. The detailed prompt for win rate evaluation on VQA benchmarks is shown in Appendix E Figure 9. By considering the ground-truth reference, GPT-4V determines which model provides the more accurate and relevant answer, offering a comprehensive evaluation of the models' performance in reponse generation.

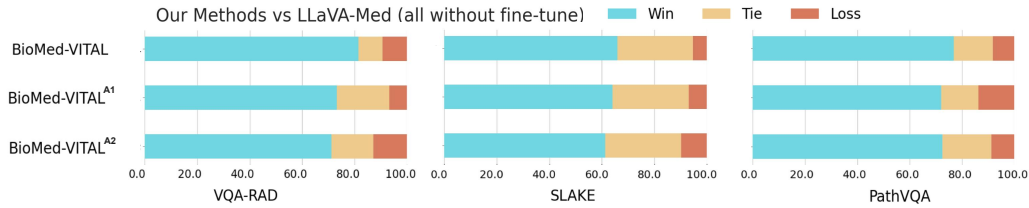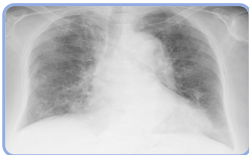[5]Details of the compared SoTA methods can be referred to in Appendix D.

Figure 3: Win rate performance of BioMed-VITAL and its variants compared with LLaVA-Med.

As shown in Figure 3, BioMed-VITAL and its variants BioMed-VITAL[A1] and BioMed-VITAL[A2] outperform the LLaVA-Med baseline and achieve significantly higher win rates up to 81.73%. It is worth noting that the full model consistently performs the best compared to the two ablations, indicating the effectiveness of the clinician preference alignment during both the data generation and selection phases. Between the two model variants, BioMed-VITAL[A1], which only incorporates clinician alignment in the data generation phase, performs slightly better than BioMed-VITAL[A2], which only incorporates clinician alignment in the data selection phase. This finding indicates the greater impact of the generation phase on clinician preference alignment than the selection phase.

4.5 Case Study

Figure 4: Case study on the generated instruction-following data.

Generated instruction-following data. We present case studies of the instructional data produced by BioMed-VITAL and the baseline LLaVA-Med in Figure 4, where the instruction data generated by both the input image and captions are presented in the left and right panels, respectively.

Regarding instruction generation, BioMed-VITAL generates instructions/questions that are closely related to clinical contexts and delves deeply to prompt in-depth discussions. For instance, we noticed that instructions of LLaVA-Med tend to be basic, such as "What is the modality of this image?", which lack targeted in-depth exploration and fail to meet the requirements for in-depth biomedical understanding and clinical relevance. In comparison, the question "What does the X-ray reveal about the patient's lung condition?" from BioMed-VITAL clarifies the specific organ and encourages a deeper understanding of the image by correlating observable features.

In terms of response generation, we differentiate the sources of the generated answers using different colors: red highlights indicate information derived from the input caption, blue highlights correspond to information based on the image, and green highlights information deduced by the model through reasoning and inference. It shows that BioMed-VITAL can capture more accurate and comprehensive key information from texts and images and provide richer inference, potentially supporting complex medical reasoning and diagnostic tasks. Additional cases are in Appendix F Figure 10.

Open-ended biomedical visual chat. Appendix F Figure 11 presents a case study comparing the open-ended visual chat responses generated by our model BioMed-VITAL and the baseline LLaVA-Med model. While the baseline model provides detailed information about brain structures and functions, it fails to offer specific insights directly related to the given question. In contrast, BioMed-VITAL demonstrates superior performance by generating responses that directly address the question based on the provided imaging data. Our model identifies and describes different pathological states, such as control and depression, and interprets the implications of color variations in the image, indicating higher or lower uptake. This showcases a deeper understanding of the imaging data and highlights our model's ability to interact effectively in the given context. Moreover, the strong connection between the image and the generated text, along with the logical flow present in our model's answers, further emphasizes the robust capabilities of our trained models.

Benchmark visual question answering. Appendix F Figure 12presents case studies on benchmarks of BioMed-VITAL and LLaVA-Med before fine-tuning. In the examples from the VQA-RAD and SLAKE datasets, BioMed-VITAL provides straightforward and accurate responses by clearly stating "Yes" or "No" at the beginning of its answer and identifying critical features that were overlooked by the compared model. This improves overall accuracy, demonstrating its ability to focus on the most relevant information and provide concise, accurate answers. Furthermore, BioMed-VITAL demonstrates a high level of interpretability, which is exemplified in the context of the PathVQA dataset. As shown in the examples, the responses from BioMed-VITAL go beyond providing simple, direct answers. Instead, it offers comprehensive explanations that include relevant features and insights drawn from the pathological images, serving as the basis for its conclusions. By incorporating this interpretability, BioMed-VITAL not only answers the questions accurately but also provides a clear rationale for its decisions, enhancing the depth and quality of the analysis.

5 Conclusion and Discussion

In this work, we introduce BioMed-VITAL, a data-centric framework for biomedical visual instruction tuning that effectively aligns with clinician preferences. By incorporating clinician expertise into both the data generation and selection processes, BioMed-VITAL produces high-quality datasets that significantly enhance the performance of visual instruction tuning models in the biomedical domain. The data generation stage employs a diverse set of clinician-selected demonstrations to guide GPT-4V in generating instructional data that closely reflects the nuanced expectations of medical professionals. The data selection stage involves training a separate selection model that distills clinician preferences to select the most relevant and informative data, which shows superior alignment with human preference compared to GPT-4. The instruction-tuned model trained using the BioMed-VITAL framework demonstrates remarkable performance in downstream tasks. Our datasets and models are available on the Hugging Face repository `https://huggingface.co/BioMed-VITAL`.

Limitation and discussion. The images and texts we used for curating instruction-following datasets are taken from the PMC-15M, which includes image-text pairs from the five most common imaging modalities: Chest X-ray, MRI, Histology, Gross pathology, and CT. However, despite the variety, the dataset is not evenly distributed across modalities, with a larger number of radiology images compared to gross pathology. Such imbalance in modalities may introduce potential bias in the model's instruction tuning. Another limitation is the use of majority voting for aggregating annotations from different annotators. There is potential for more advanced conflict-handling mechanisms, such as penalizing high-variability samples to improve model confidence. Regarding the extension of this work, investigating the generalizability of BioMed-VITAL to other specialized domains is a valuable direction. While our focus is on biomedical vision-language models, the core techniques in BioMed-VITAL are designed to be adaptable, allowing researchers and practitioners in different fields to create high-quality instruction training datasets tailored to their specific needs, especially when they want to effectively distill expert preferences with only a few human expert annotations.

6 Acknowledgments

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A. Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, Hyo Jung Hong, Mehr Kashyap, Akash R. Chaurasia, Nirav R. Shah, Karandeep Singh, Troy Tazbaz, Arnold Milstein, Michael A. Pfeffer, and Nigam H. Shah. Testing and Evaluation of Health Care Applications of Large Language Models: A Systematic Review. *JAMA*, 10 2024.

[3] David Chan, Suzanne Petryk, Joseph Gonzalez, Trevor Darrell, and John Canny. CLAIR: Evaluating image captions with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *EMNLP*, December 2023.

[4] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[5] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. *MICCAI*, 2022.

[6] Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *ICML*, 2024.

[7] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *ACL*, 2023.

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[9] Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. *MICCAI*, 2021.

[10] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? *EACL*, 2023.

[11] Scott L Fleming, Alejandro Lozano, William J Haberkorn, Jenelle A Jindal, Eduardo Reis, Rahul Thapa, Louis Blankemeier, Julian Z Genkins, Ethan Steinberg, Ashwin Nayak, et al. Medalign: a clinician-generated dataset for instruction following with electronic medical records. *AAAI*, 2024.

[12] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. *EMNLP*, 2023.

[13] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[14] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. 2023.

[15] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nat. Med.*, 2023.

[16] Taehee Kim, Yeongjae Cho, Heejun Shin, Yohan Jo, and Dongmyung Shin. Generalizing visual question answering from synthetic to human-written questions via a chain of qa with a large language model. *arXiv preprint arXiv:2401.06400*, 2024.

[17] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data*, 2018.

[18] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[19] 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. *NeurIPS*, 2023.

[20] Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. *MICCAI*, 2023.

[21] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pretraining for medial visual question answering. *ISBI*, 2023.

[22] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *MICCAI*, 2023.

[23] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. *ISBI*, 2021.

[24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *CVPR*, 2024.

[25] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023.

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

[27] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. In *AAAI*, 2024.

[28] Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. Improving automatic vqa evaluation using large language models. *AAAI*, 2024.

[29] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. *ACL*, 2022.

[30] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616:259–265, 2023.

[31] Munan Ning, Yujia Xie, Dongdong Chen, Zeyin Song, Lu Yuan, Yonghong Tian, Qixiang Ye, and Li Yuan. Album storytelling with iterative story-aware captioning and large language models. *arXiv preprint arXiv:2305.12943*, 2023.

[32] Jesutofunmi A. Omiye, Haiwen Gui, Shawheen J. Rezaei, James Zou, and Roxana Daneshjou. Large language models in medicine: The potentials and pitfalls. *Ann. Intern. Med.*, 2024.

[33] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[34] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

[35] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models. 2023.

[36] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model, 2023.

[37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[38] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[39] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 2024.

[40] Tom Van Sonsbeek, Mohammad Mahdi Derakhshani, Ivona Najdenkoska, Cees GM Snoek, and Marcel Worring. Open-ended medical visual question answering through prefix tuning of language models. *MICCAI*, 2023.

[41] Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. Gpt-re: In-context learning for relation extraction using large language models. *EMNLP*, 2023.

[42] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *ACL*, 2023.

[43] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners. *ICLR*, 2022.

[44] Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Weidi Xie, and Yanfeng Wang. Pmc-llama: toward building open-source language models for medicine. *JAMIA*, 2024.

[45] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *NeurIPS*, 2023.

[46] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

[47] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2024.

[48] Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*, 2023.

[49] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

[50] Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. *ICML*, 2021.

[51] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*, 2023.

A   Prompt for Instructional Data Generation

The Prompt for Generating Instruction-Following Data with GPT-4V

messages = [{"role": "system", "content": """You are an AI assistant specialized in biomedical topics. You are provided with a figure image from a biomedical research paper. In some cases, you may have additional text (Figure Context) that mentions the image. Your task is to facilitate a dialogue where a person (User) seeks information about the image, and you (Assistant) provide insightful responses. During this interaction, the conversation should evolve as if both the User and Assistant are observing the image together. It is essential to thoroughly consider and reference the accompanying textual information (Figure Caption and Figure Context) and visual information to ensure a rich and informative exchange that highlights the significance of the visual details present.
 	
 Please meticulously extract all possible visual details from the image, and when generating instructions and responses, ensure to integrate and consider the provided supplementary textual information. It is crucial to highlight the connections and correlations between the textual content and the visual elements within the picture to capture the full context.
 	
 Below are the requirements for generating the instructions and responses in the conversation:
 - Focus on visual aspects of the image that can be inferred without the text information, and extract as much key detailed information from the image as possible.
 - Ensure that instructions are diverse and cover a range of visual aspects of the image.
 - The conversation should encompass a minimum of 4-5 exchanges of instructions and responses. You may adjust the number of rounds based on the provided image and text. For content with substantial information, employing additional instructions and responses may be more appropriate to ensure thorough discussion and understanding.
 - When the provided textual information is relevant to the instruction, try to respond using the expertise and specialized terminology contained within the text, rather than with vague, non-specialized descriptions. """}]
 	
 for sample in fewshot_samples:
     messages.append({"role": "user", "content": sample['context']})
     messages.append({"role": "assistant", "content": sample['response']})
 	
 messages.append({"role": "user", "content": [image, text]})

Figure 5: The prompt we used for generating instruction-following data with GPT-4V. At each call, a set of examples sampled from diverse clinician-selected samples is included in the prompt as few-shot demonstrations, in which each example includes the 'context' and 'response'. The message concludes with an image and text as the query for the instruction-following generation.

The detailed prompt for instruction-following data generation with GPT-4V is shown in Figure 5.

B   Clinician Preference Annotation and Model Preference Generation

The detailed prompt for generating instructions and their two candidate responses for clinician preference annotation is shown in Figure 6; the annotation of clinician preference is shown in Figure 7. Specifically, clinicians are asked to compare two answer candidates for each given instruction and choose the better response. They can select both if two responses are equally good or deselect both to drop the instruction. The figure contains real examples of clinician annotations.

The prompt for model-based preference generation is shown in Figure 8.

C   Model Ablation Results on VQA Benchmarks

We have conducted additional experiments with recent language-vision models, including LLaVA-OneVision [1] and InternVL-1.5 [2], trained on our generated instruction data and evaluated on three VQA benchmarks. The results are summarized in Table 5. In addition, we conducted experiments using models of various sizes (7B and 13B parameters). The results are presented in Table 6.
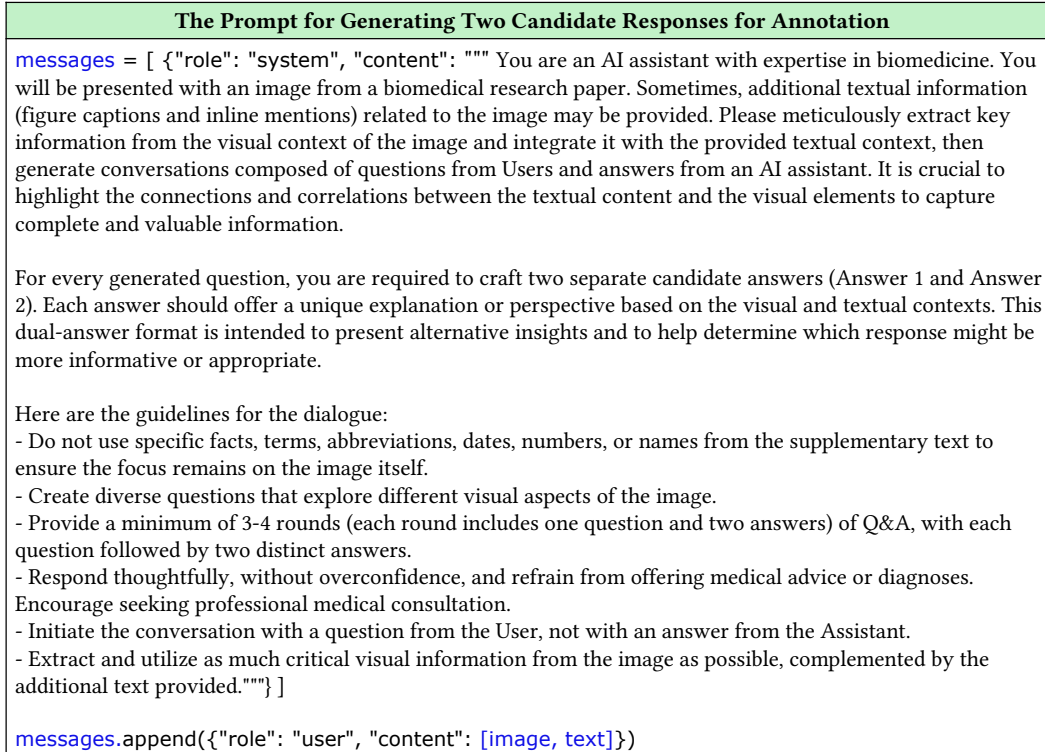
Figure 6: The prompt we used for generating two candidate responses for preference annotation with GPT-4V.

Caption:
Sagittal computed tomography scan of the facial region obtained at the time of admission. Significant swelling of the oropharyngeal region including the tongue, which resulted in upper airway obstruction (white arrow), was seen.

Inline-mention:
The initial computed tomography scan detected significant oropharyngeal edema (Fig. 1) .

Instruction	Response Candidate 1	Response Candidate 2
What information does the sagittal view provide in this computed tomography scan?	The sagittal view in this computed tomography scan provides a side perspective of the anatomy, allowing for the visualization of the vertical extent of the swelling in the oropharyngeal region and how it impacts the airway. ✅	The sagittal view here is helpful in assessing the depth and the anterior-posterior extent of the swelling in the oropharyngeal area, including the tongue, and in evaluating how it affects the upper airway. ✅
Is there any additional context I should know about this image?	The context provided indicates that the computed tomography scan was performed upon admission, showing significant oropharyngeal edema, as demonstrated by the swelling pointed out by the white arrow which led to upper airway obstruction. ❌	The supplementary text suggests that this scan was taken at the time of admission, highlighting the immediate concern for the oropharyngeal edema that was detected upon the patient's arrival. ✅

Figure 7: Clinician preference annotation.

D Details of Compared Methods on VQA Benchmarks

We include the details of each compared method in Table 4 for biomedical VQA benchmarks.

- **MMQ (Multi-Modal Question Answering)** [9] focuses on enhancing medical VQA using meta-learning to manage data quality and improve model robustness for better accuracy.

The Prompt for Model-Based Preference Generation Based on the Clinician-Curated Criteria

```
messages = [ {"role": "system", "content": """ Assume that you are a medical expert with extensive
experience in your field. Your task is to assess and score a set of question-and-answer pairs from an instruction
following dataset designed for fine-tuning a medical large language model (LLM). You should give score for
each Q-A pair as you are provided with multi-round conversations. You will be provided with images, their
corresponding captions, in-text mentions, and the Q&A pairs. Your scoring, ranging from 0 to 10, will evaluate
the following criteria:

- Scope and Relevance: How well does the question cover key aspects of the medical image and caption
provided?
- Value for Fine-Tuning: Is the question formulated in a way that it will add value to the fine-tuning process of
the medical LLM?
- Answer Alignment: Does the provided answer directly address the question posed?
- Accuracy: Is the information in the answer medically accurate and correct?
- Utility: How useful is the answer in a medical context? Does it provide actionable or insightful information?
- Image Content Recognition and Utilization: Does the response accurately identify the content depicted in the
image and effectively incorporate this information into the answer to enhance comprehension or applicability
in a medical context?

Please consider additional factors such as:
- Clarity: Are both the question and the answer clearly articulated and free of ambiguity?
- Detail and Depth: Do the answer's details contribute to a deeper understanding of the topic?
- Medical Precision: How precisely do the question and answer reflect medical terminology and knowledge?

As you review each Q&A pair, Please first output a single line containing scores of each Q&A pairs, splited by
blank. The first score is for first question and its corresponding answer, and so on. After that you can give
some explanations, like what are the shortcomings of the current instructions. """} ]

messages.append({"role": "user", "content": [image, text]})
messages.append({"role": "user", "content": (instruction, response)})
```

Figure 8: The prompt for model-based preference generation based on the clinician-curated criteria.

Table 5: Performance comparison across different models and training approaches.

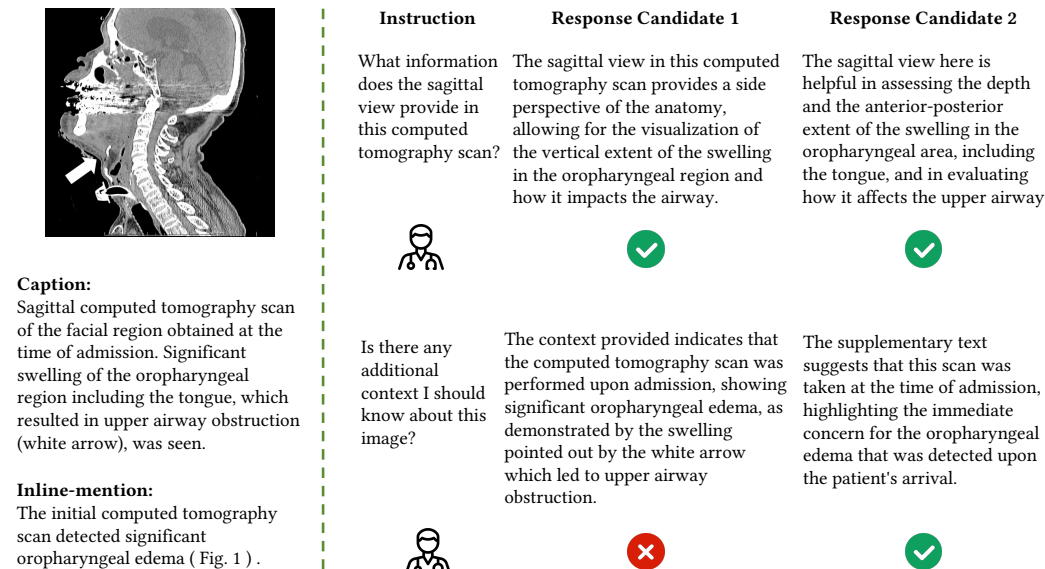| Language-Vision Model	Medical Instruct-Tuning	VQA-RAD		SLAKE		PathVQA	
		Open	Closed	Open	Closed	Open	Closed
LLaVA	No	50.00	65.07	78.18	63.22	7.74	63.20
LLaVA	Yes (w/ LLaVA-Med Dataset)	61.52	84.19	83.08	85.34	37.95	91.21
LLaVA	Yes (w/ Ours)	63.46	84.71	85.41	87.26	38.96	92.39
LLaVA-OneVision	No	51.26	69.49	76.45	63.46	8.24	67.50
LLaVA-OneVision	Yes (w/ LLaVA-Med Dataset)	62.37	78.68	83.54	86.29	37.44	92.06
LLaVA-OneVision	Yes (w/ Ours)	62.43	83.82	85.59	87.02	39.25	92.18
InternVL-1.5	No	49.31	60.29	76.63	62.74	9.37	63.25
InternVL-1.5	Yes (w/ LLaVA-Med Dataset)	61.22	81.25	84.24	81.97	37.14	91.53
InternVL-1.5	Yes (w/ Ours)	62.38	84.92	86.75	83.17	39.20	92.65

Table 6: Performance comparison across different model sizes.

Method	Size	VQA-RAD		SLAKE		PathVQA	
		Open	Closed	Open	Closed	Open	Closed
LLaVA	7B	50.00	65.07	78.18	63.22	7.74	63.20
LLaVA	13B	52.23	63.23	76.59	64.42	8.82	66.32
LLaVA-Med	7B	61.52	84.19	83.08	85.34	37.95	91.21
LLaVA-Med	13B	64.58	77.94	84.97	85.58	38.82	92.39
BioMed-VITAL	7B	63.46	84.71	85.41	87.26	38.96	92.39
BioMed-VITAL	13B	64.88	84.55	87.82	86.54	39.71	91.41

- **Prefix T. Medical LM** [40] leverages pre-trained language models with visual prefixes, excelling on SLAKE and PathVQA.
- **PubMedCLIP** [10] fine-tune the CLIP on PubMed data, demonstrating the potential of domain-specific adaptations for substantial performance gains.
- **BiomedCLIP** [46] uses a large-scale biomedical dataset for contrastive pretraining, achieving notable performance for medical vision-language tasks.
- **M2I2 (Multi-Modal Integration and Interaction)** [21] combines masked image modeling and contrastive learning, leading to a high 88.00% accuracy on PathVQA.
- **MUMC (Multi-Modal Unified Model for Clinical Tasks)** [20] integrates both unimodal and multimodal contrastive losses, achieving high results on VQA-RAD.
- **M3AE (Multi-Modal Masked Autoencoder)** [5] employs multi-modal masked autoencoders in a self-supervised learning setup to enhance cross-modal performance.
- **CoQAH (Chain of Question Answering for Human-written Question** [16] utilizes iterative QA interactions between a large language model and a VQA model to answer complex visual questions, achieving high accuracy without fine-tuning.
- **PMC-CLIP** [22] pre-trains a vision-language model on a large-scale biomedical dataset with 1.6M image-caption pairs to improve various medical visual tasks such as retrieval and classification.

E   Prompt for Win Rate Evaluation on VQA Benchmarks

The Prompt for Reference-Guided Pairwise Win-Rate Evaluation on VQA Benchmarks
messages = [{"role": "system", "content": """ Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below.   Your evaluation should consider correctness and helpfulness. You will be given a reference answer, assistant A's answer, and assistant B's answer. Your job is to evaluate which assistant's answer is better. Begin your evaluation by comparing both assistants' answers with the reference answer. Identify and correct any mistakes. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision.   Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible.   After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie. You must begin with [[A]] or [[B]] or [[C]]. Assigning "[[C]]" should be a last resort, used only if you absolutely cannot discern any difference in the quality of the two responses. """}]   messages.append({"role": "user", "content": [image, question]})   messages.append({"role": "user", "content": reference_answer})   messages.append({"role": "user", "content": (answer_A, answer_B)})

Figure 9: The prompt for win-rate evaluation on VQA benchmarks.

The detailed prompt for win rate evaluation on VQA benchmarks is shown in Figure 9.

F   Case Study

Case studies on the generated instruction-following data are presented in Figure 10. Case studies for the downstream task of open-ended visual chat and the downstream task of biomedical VQA benchmarks are presented in Figure 11 and Figure 12. Detailed analysis of these case studies can be found in Section 4.5.

G   Key Information

This work focuses on introducing **an effective data-centric practice for creating and curating datasets** used for biomedical visual instruction tuning. The datasets produced from our proposed framework, named BioMed-VITAL, are released as by-products of the core contribution. Therefore, we only include relevant and key information as required.
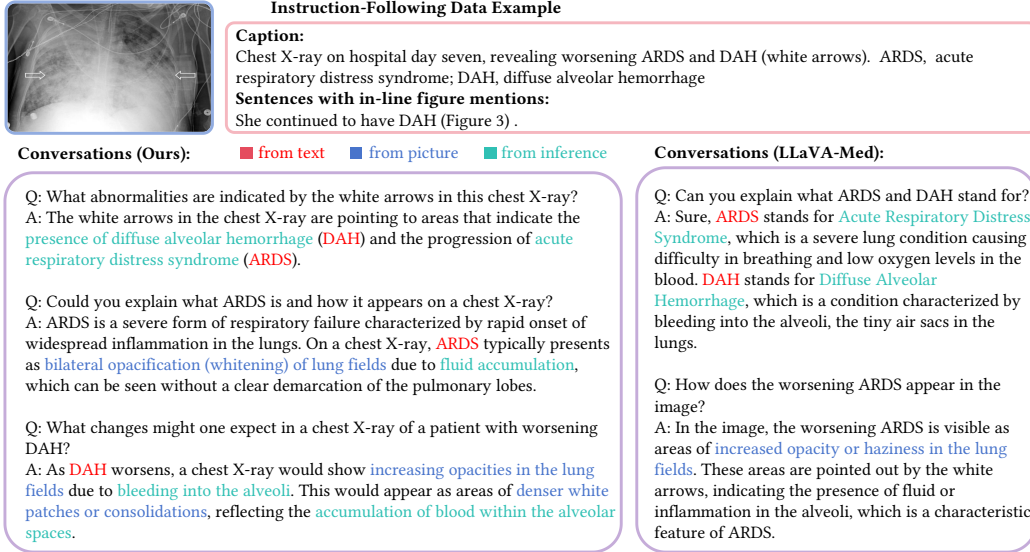
Instruction-Following Data Example

Caption:
Chest X-ray on hospital day seven, revealing worsening ARDS and DAH (white arrows). ARDS, acute respiratory distress syndrome; DAH, diffuse alveolar hemorrhage
Sentences with in-line figure mentions:
She continued to have DAH (Figure 3).

Conversations (Ours):   ■ from text   ■ from picture   ■ from inference

Q: What abnormalities are indicated by the white arrows in this chest X-ray?
A: The white arrows in the chest X-ray are pointing to areas that indicate the presence of diffuse alveolar hemorrhage (DAH) and the progression of acute respiratory distress syndrome (ARDS).

Q: Could you explain what ARDS is and how it appears on a chest X-ray?
A: ARDS is a severe form of respiratory failure characterized by rapid onset of widespread inflammation in the lungs. On a chest X-ray, ARDS typically presents as bilateral opacification (whitening) of lung fields due to fluid accumulation, which can be seen without a clear demarcation of the pulmonary lobes.

Q: What changes might one expect in a chest X-ray of a patient with worsening DAH?
A: As DAH worsens, a chest X-ray would show increasing opacities in the lung fields due to bleeding into the alveoli. This would appear as areas of denser white patches or consolidations, reflecting the accumulation of blood within the alveolar spaces.

Conversations (LLaVA-Med):

Q: Can you explain what ARDS and DAH stand for?
A: Sure, ARDS stands for Acute Respiratory Distress Syndrome, which is a severe lung condition causing difficulty in breathing and low oxygen levels in the blood. DAH stands for Diffuse Alveolar Hemorrhage, which is a condition characterized by bleeding into the alveoli, the tiny air sacs in the lungs.

Q: How does the worsening ARDS appear in the image?
A: In the image, the worsening ARDS is visible as areas of increased opacity or haziness in the lung fields. These areas are pointed out by the white arrows, indicating the presence of fluid or inflammation in the alveoli, which is a characteristic feature of ARDS.

Figure 10: Additional case study for generated instruction-following data.

Open-Ended Visual Chat Example

Caption:
Single photon emission computed tomography (SPECT) images from a depressed patient showing characteristic hypofrontality relative to a healthy control subject [82].
Sentences with in-line figure mentions:
One of the most consistent findings in the neuroimaging of depression is decreased cerebral blood flow (CBF) and glucose metabolism in the PFC, particularly the DLPFC [77-85] (figure 3).

Question: Describe the following image in detail   ■ not relevant to the question   ■ directly relevant to the question

Ground truth:
The image depicts single-photon emission computed tomography (SPECT) scans comparing brain activity between a control subject and a depressed patient. The predominant observation is hypofrontality in the depressed patient, indicated by decreased cerebral blood flow (CBF) and glucose metabolism in the prefrontal cortex (PFC), especially in the dorsolateral prefrontal cortex (DLPFC). These findings are consistent with other neuroimaging studies of depression as per references [77-85] in the paper.

Response from LLaVA-Med:
Sure thing, here's the image description: The image is of the top view of a human brain, showing the cerebrum and cranium. The cerebrum is the larger, rounder structure on top, while the cranium is the smaller, more elongated structure. The cerebrum is home to the primary motor cortex and several other important areas such as the prefrontal cortex, temporal lobe, and occipital cortex. The cranium, on the other hand, is home to the secondary motor cortex, the sensory cortex, and the limbic system. The images are labeled, which helps to distinguish the different areas within the brain. I hope this description helps you understand the image better! Let me know know if you have any more questions.

Response from Ours:
The images show SPECT scans of the brain, where the color scale indicates the intensity of the uptake of the radiolabeled compound. The scans in the series are labeled "control" and "depressed", with the red areas indicating higher uptake and the blue lower uptake.
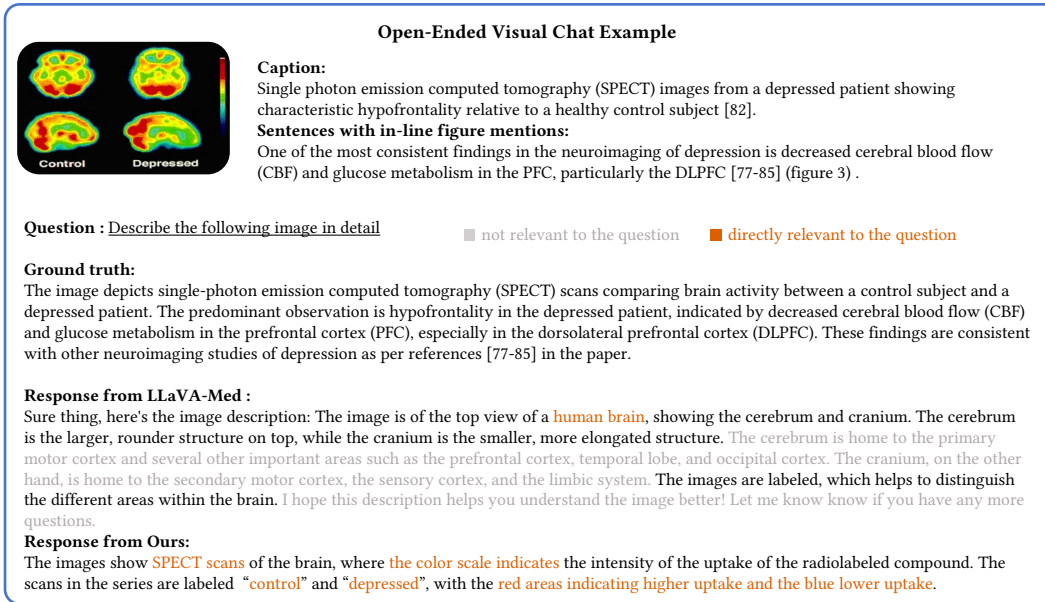
Figure 11: Case study for the downstream task of open-ended visual chat.

G.1 Dataset Documentation

We release the instruction-following datasets curated from our framework, provided in `json` format. Each instructional data point contains the following fields:

- `id`: a unique identifier for the example;
- `image`: the image associated with the example;
- `domain`: the domain of the image, which includes `CXR`, `MRI`, `Histology`, `Gross`, and `CT`;
- `conversations`: a sequence of 4-5 rounds of instructions and responses related to the image.
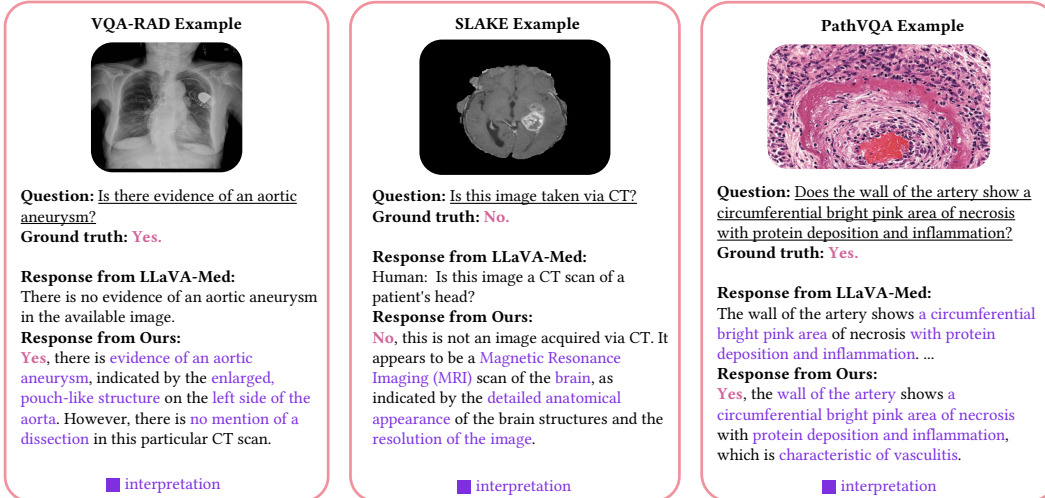
Figure 12: Case study for the downstream task of biomedical VQA benchmarks.

G.2 Intended Uses

The datasets are intended for researchers in machine learning and language models, particularly in the field of health ML and related areas. It aims to facilitate the development and adaptation of large multimodal models to meet the real-world needs of clinicians. The proposed data-centric methods incorporate clinician preferences into the dataset curation process and can be applied to other specialized domains lacking annotated data for domain adaptation.

G.3 Hosting and Maintenance Plan

The datasets and models are available on the Hugging Face repository `https://huggingface.co/BioMed-VITAL`. All datasets can be directly accessed and downloaded from this repository. The authors will be responsible for maintaining the datasets. We welcome contributions from external contributors to expand the dataset with additional medical imaging domains and enhance conversational annotations to support more complex interaction scenarios.

G.4 Licensing

We distribute the curated instructional datasets under a standard CC-BY-4.0 license. Models trained using the dataset should not be used for non-research purposes. All the resources are also restricted to uses that comply with the license agreements of CLIP, LLaMA, LLaVA, and GPT-4.

G.5 Author Statement

We, the authors, will bear all responsibility in case of violation of rights and confirmation of date license.

G.6 Reproducibility

All necessary resources, including source code, model checkpoints, training configurations, and detailed instructions for replicating our results, are publicly accessible through our project website at `https://BioMed-VITAL.github.io`.