MedMax: Mixed-Modal Instruction Tuning for Training Biomedical Assistants

Hritik Bansal **Daniel Israel*** **Siyan Zhao*** **Shufan Li** **Tung Nguyen**
Aditya Grover

University of California Los Angeles

December 18, 2024

Abstract

Recent advancements in mixed-modal generative models have enabled flexible integration of information across image-text content. These models have opened new avenues for developing unified biomedical assistants capable of analyzing biomedical images, answering complex questions about them, and predicting the impact of medical procedures on a patient's health. However, existing resources face challenges such as limited data availability, narrow domain coverage, and restricted sources (e.g., medical papers). To address these gaps, we present MedMax, the first large-scale multimodal biomedical instruction-tuning dataset for mixed-modal foundation models. With 1.47 million instances, MedMax encompasses a diverse range of tasks, including multimodal content generation (interleaved image-text data), biomedical image captioning and generation, visual chatting, and report understanding. These tasks span diverse medical domains such as radiology and histopathology. Subsequently, we fine-tune a mixed-modal foundation model on the MedMax dataset, achieving significant performance improvements: a 26% gain over the Chameleon model and an 18.3% improvement over GPT-4o across 12 downstream biomedical visual question-answering tasks. Additionally, we introduce a unified evaluation suite for biomedical tasks, providing a robust framework to guide the development of next-generation mixed-modal biomedical AI assistants.

Code: https://github.com/Hritikbansal/medmax
Website: https://mint-medmax.github.io/

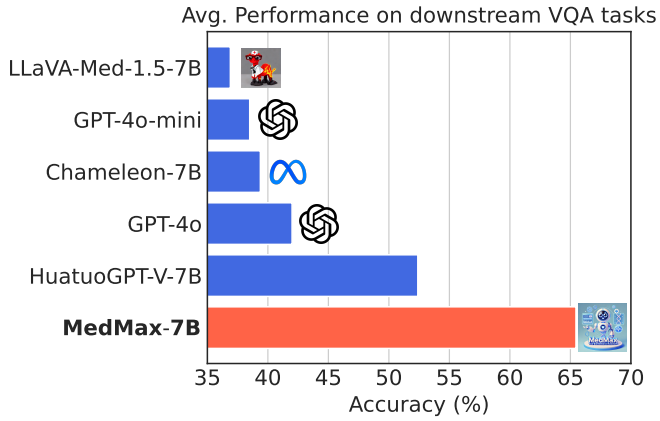


Figure 1: **Average performance of multimodal models on twelve VQA tasks.** Our MedMax instruction-tuned mixed-modal foundation model outperforms both open multimodal models (Chameleon, LLaVA-Med-v1.5, and Huatuo) and closed multimodal models (GPT-4o, GPT-4o-mini). This underscores the effectiveness of the MedMax dataset in training capable multimodal biomedical assistants.

∗ Equal Contribution (alphabetical order).

1 Introduction

Recently, there has been rapid advancement in the development of mixed-modal foundation models that can perceive and generate data from multiple modalities such as Chameleon [54], Transfusion [76], Aurora [8], Gemini-2.0-Flash [18], and others [75, 58, 51, 70]. In particular, these models offer a flexible design, which embeds multiple modalities into a shared embedding space and uses a transformer backbone to learn interactions between these diverse modalities. During pretraining, these models are exposed to internet-scale data that equips them with the knowledge to perform real-world tasks involving multiple modalities within a unified architecture (e.g., image captioning and image generation). The ability of foundation models to understand and reason across the vast landscape of biomedical data—including multiple modalities (e.g., image-caption pairs), sources (e.g., medical research papers, YouTube channels), and domains (e.g., radiology, histology)—positions them as powerful tools for biomedical diagnosis, prognosis, and prevention [6, 59]. their multimodal generation capabilities can facilitate medical treatment planning by predicting a patient's CT scan outcomes following specific treatments.

However, existing mixed-modal foundation models struggle to perform well on vision-language biomedical data (e.g., visual question answering on X-rays) due to significant distribution shifts from the more commonly occurring natural data found on the internet (e.g., everyday objects and scenes). In this context, instruction tuning [60, 39, 29] offers a promising approach to understanding novel user intents and unlocking new capabilities for developing advanced biomedical assistants. But, there is a lack of large-scale multimodal biomedical instruction-tuning datasets, which are crucial for enabling mixed-modal models to reason and solve complex biomedical tasks across diverse domains.

Traditional biomedical visual question answering (VQA) datasets such as VQA-RAD [30], SLAKE [36], and PathVQA [20] are crucial for imparting domain-specific knowledge. However, they suffer from several limitations such as lack of scale (e.g., typically only thousands of instances), task diversity (e.g., not suitable for teaching biomedical captioning or generation), and narrow focus on specific domains (e.g., radiology or pathology). In contrast, LLaVA-Med [39] collects biomedical vision-language alignment data (i.e., image-text pairs) and curated synthetic instruction-tuning datasets (i.e., image-conditioned conversational data). This aids in allowing the foundation models to assist the practitioners for novel queries about the biomedical images. However, it relies on PMC-15M [72], which is not openly available. Moreover, much of the LLaVA-Med data consists of figures and plots rather than biomedical images, impacting its overall quality.

Subsequent work like PubMedVision [71] aims to improve data quality through synthetic data curation across diverse biomedical domains. Nevertheless, it is limited to the knowledge available in medical research papers, restricting its ability to generate medical reports (e.g., findings and impressions based on biomedical imagery) that could assist doctors. Other datasets, such as Quilt-LLaVA, independently create visual instruction-tuning data for histopathology using alternative data sources, such as YouTube videos. Similarly, MIMIC-CXR provides expert-written radiology reports for chest X-rays. Furthermore, the ability to generate biomedical images is crucial for alleviate the lack of annotated medical imaging datasets [13] and disease progression modeling [19]. However, there are no existing biomedical instruction tuning datasets that focus on biomedical text-to-image generation and interleaved image-text content.

To address these challenges, we propose MedMax, a dataset designed to develop a biomedical mixed-modal foundation model. It comprises a total of 1.47M instances spanning a wide range of biomedical tasks and domains. Specifically, MedMax includes tasks such as biomedical image captioning, image generation, visual question answering (VQA), visual chatting, report understanding, and multimodal (interleaved text-image) content generation. Moreover, the dataset encompasses diverse biomedical domains, including radiology and histopathology. A key component of MedMax is a newly curated dataset for generating interleaved image-text content (MedMax-Instruct), which paves the way for enhanced clinical understanding and support for complex medical decision-making. Additionally, MedMax aims to equip mixed-modal models with a diverse skill set by integrating various high-quality multimodal datasets, including VQA datasets, instruction-following datasets, alignment datasets, and medical reports.

Subsequently, we fine-tune a mixed-modal foundation model, Chameleon [54, 15], on the MedMax dataset. The dataset comprises a total of 1.7B multimodal discrete tokens for instruction tuning. In our experiments, the MedMax-fine-tuned model outperforms the base Chameleon and GPT-4o [23] by 26% and

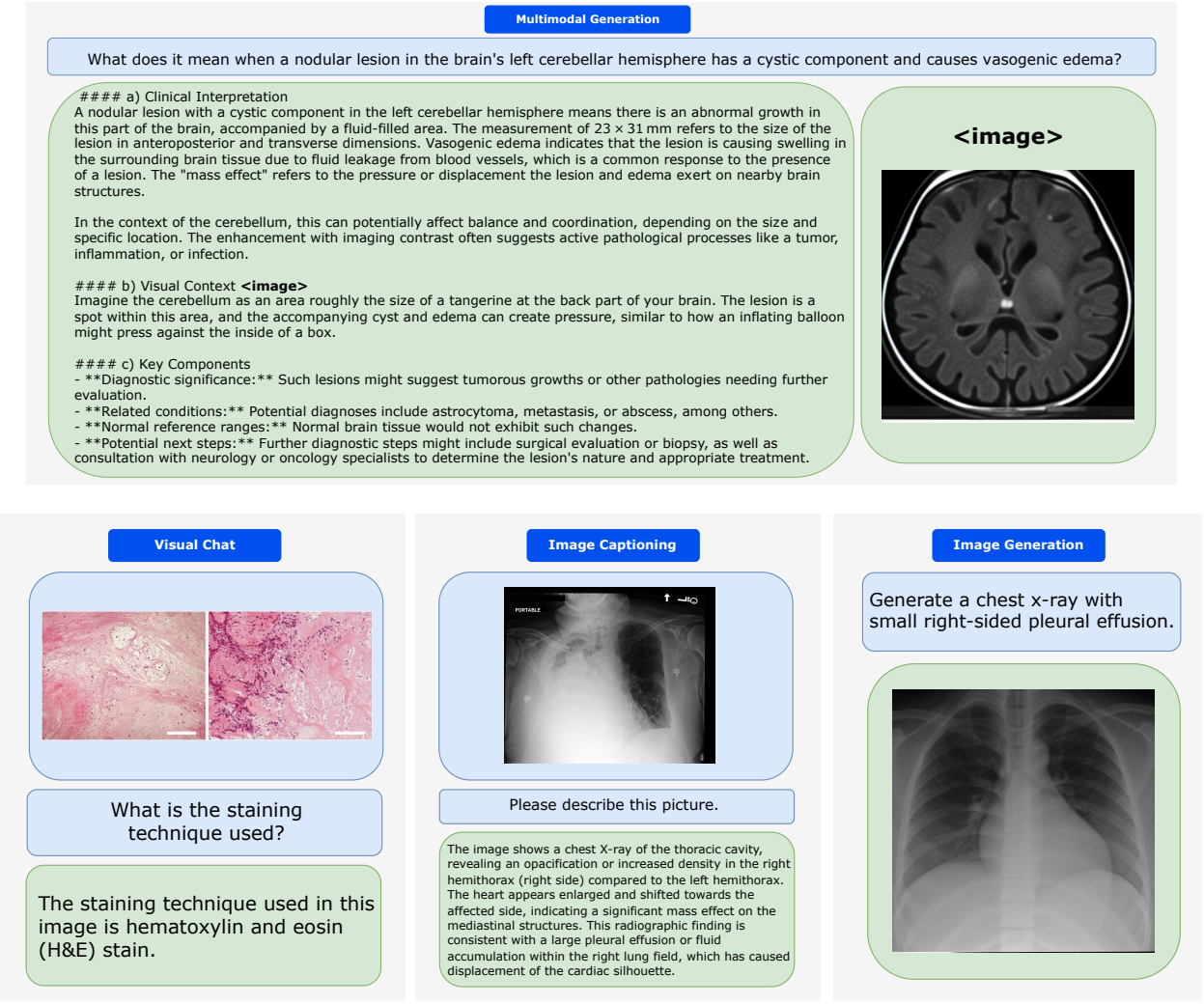


Figure 2: **Qualitative Examples.** Sample generations from the MedMax model across diverse tasks: multimodal (interleaved) generation, visual chat, image captioning, and image generation.

18.3% percentage points, respectively, when averaged across a set of 12 downstream VQA evaluation tasks (Figure 1). Furthermore, we conduct extensive experiments on diverse tasks that mixed-modal models excel in, including biomedical image captioning, image generation, visual chatting, and multimodal data generation. Given the general lack of support for biomedical evaluation, we provide a comprehensive evaluation suite encompassing diverse tasks to enable unified and efficient assessments. Overall, our work establishes a strong foundation for high-quality instruction tuning data creation, model fine-tuning, and robust evaluation of next-generation mixed-modal models.

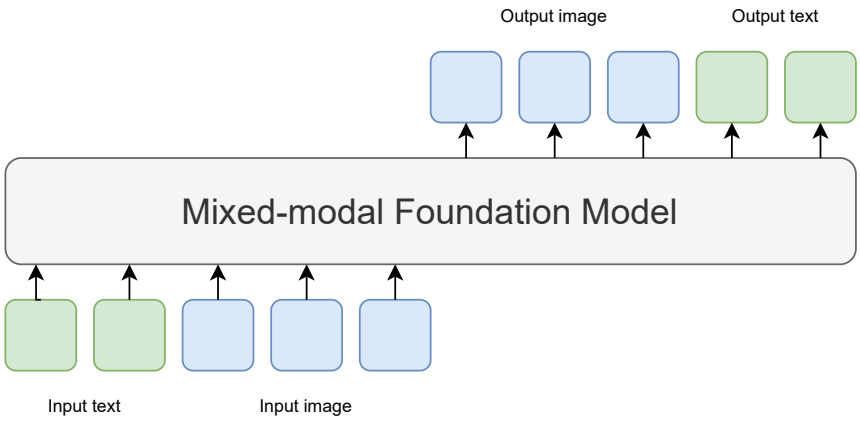


Figure 3: A mixed-modal foundation model is capable of understanding text and image inputs and can generate both textual and visual outputs through a unified architecture.

2 MedMax

3 Background

3.1 Mixed-Modal Foundation Models

Mixed-modal foundation models are a class of generative models that can reason over the sequence of arbitrarily interleaved multimodal (e.g., image, text) content [54, 76]. In this work, we primarily focus on the autoregressive sequence modeling objective, as used by Chameleon [54], Unified-IO [40], and Emu-3 [58], for its simplicity. Formally, these methods model interleaved multi-modal tokens $x = (x_1, x_2, \ldots, x_n)$ where n is the length of the sequence. Here, the text content is represented as BPE tokens, while image tokens are obtained from an image encoder. For instance, Chameleon uses 1024 tokens obtained from a VQGAN [17] encoder to represent each image. Given a dataset $\mathcal{D}$, an autoregressive pretraining objective for multi-modal data $x \in D$ can be formulated as:

$$\max_{\theta} \mathbb{E}_{x \sim \mathcal{D}} \left[\sum_{k=1}^{n} \log P_\theta(x_k | x_{1:k-1}) \right]. \tag{1}$$

By incorporating a diverse set of multimodal sequences at an internet scale, these pretrained models can achieve strong performance on downstream tasks such as image generation and image captioning. We illustrate the architecture for a mixed-modal foundation model in Figure 3.

3.2 Instruction Tuning

While internet-scale pretraining equips foundation models with diverse world knowledge, further instruction tuning is conducted to develop new skills and teach them how to interact with humans as assistants [60, 53]. In this context, it is essential to instruction-tune mixed-modal foundation models to impart task-specific skills. Formally, the instruction tuning data comprises paired multimodal sequences (x, y), where x represents the 'instruction' and y represents the 'response.' Both x and y may include image tokens, text tokens, or a combination of both. For instance, in the context of visual question answering (VQA) tasks, x includes an input image along with a corresponding question, while y provides the correct response. The instruction tuning objective for a dataset $\mathcal{D}_\mathcal{I}$ can be formalized as:

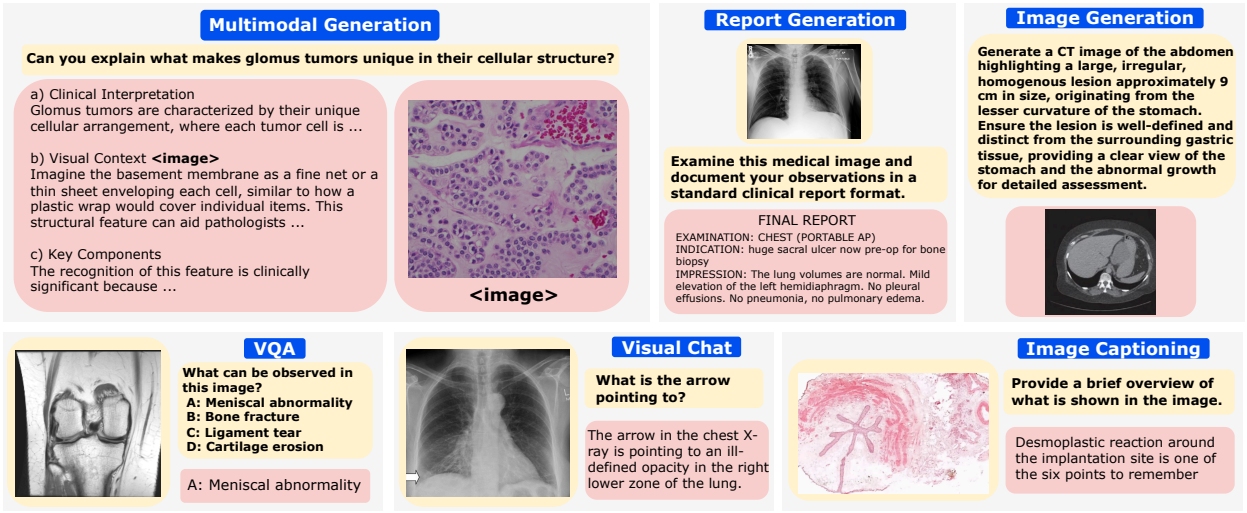


Figure 4: **Examples of diverse multimodal biomedical tasks covered in the MedMax dataset.** The model inputs (yellow boxes) and corresponding outputs (red boxes) illustrate various task types: multimodal generation with interleaved text and images, medical report generation, text-to-image generation, visual question answering, medical image analysis through visual chat, and image captioning task. Note that report-conditioned image generation, which falls under report understanding, is not shown here.

$$\max_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathcal{I}}} \left[\sum_{k=1}^{n} \log P(y_k | y_{1:k-1}, x) \right]. \tag{2}$$

Unlike the pretraining objective, which models the entire sequence, the instruction tuning objective computes the loss solely on the 'response' portion of the sequence, thereby concentrating the learning process on the desired tasks.

4 MedMax

Due to the internet-scale pretraining and flexibility of the mixed-modal foundation models, we aim to imbibe the capability to understand human intents to solve diverse biomedical tasks (e.g., VQA, generation), domains (e.g., radiology, histopathology), and modalities (e.g., image and text content) through instruction tuning. Thus, we present MedMax, an instruction tuning data designed training mixed-modal foundation models for the biomedical applications. Specifically, the dataset construction involves: (a) designing a new instruction-tuning data that allows interleaved image-text content as a model response, (b) collecting various data sources to endow diverse skills into the model, and (c) performing dataset-specific curation to ensure high-quality. We illustrate the examples for diverse skills in Figure 4. We will go through each of these components in this section.

4.1 Biomedical Multimodal Generation Instruction Tuning (MedMax-Instruct)

With the emergence of mixed-modal foundation models, generating interleaved image-text content has become possible. Specifically, such multimodal output capabilities have several novel applications, including advanced diagnostics (e.g., visualizing the effects of specific treatments on biomedical image markers), generating patient reports that are often multimodal, and enhancing medical training by providing contextually relevant multimodal content for learners. However, no instruction-tuning datasets currently exist to equip mixed-modal assistants with this capability. To address this challenge, we present a multimodal output instruction-tuning

data (MEDMAX-INSTRUCT) for training biomedical mixed-modal assistants. Broadly, this dataset is created by prompting the LLM (GPT-4o) to generate a single-turn multimodal conversation based on the image description. Prior works such as PMC-VQA [73] and LLaVA [39] have utilized the capability of the LLMs to annotate high-quality biomedical data conditioned on the variables like image description. In particular, we create this dataset in three stages:

Sourcing image-text data (Stage 1): Here, we intend to collect diverse biomedical image-text pairs that can be purposed for this task. In this regard, we combine image-caption instances from the PMC-OA and Quilt data. Specifically, our PMC-OA-specific curation (§4.2) reduces the original size of the PMC-OA (val split) from 166K to 73K instances.[1] In addition, we randomly choose 50K instances from Quilt data which does not overlap with the Quilt examples in the prior subsets of the MEDMAX data. In total, we have 123K image-text data which will be further filtered to ensure high-quality.

Filter bad captions (Stage 2): Since we aim to generate multimodal conversations conditioned on the captions, it is critical to ensure that they describe the visual contents of the image well. To address this, we filter the captions that GPT-4o-mini finds to be of low quality. We present the prompt used for assessing the caption quality in Appendix Table 4. Following this, we are left with 88K instances, leading to the removal of 25% captions. We spent $12 in performing this quality check using the API.

Caption-conditioned data generation (Stage 3): Here, we prompt GPT-4o to generate single-turn multiodal conversation conditioned on the captions in the remaining dataset. We choose GPT-4o as it achieves state-of-the-art performance on text-only medical datasets such as MedQA [43]. We present our data generation template in Appendix Table 5. Specifically, we ask the GPT model to output an image placeholder ('<image>') which is then replaced with the ground-truth image from the image-caption pair.

In total, we spent $500 to collect GPT-4o responses using the API. Finally, we have **88K** instances consisting novel query (grounded in text) and multimodal response (interleaved text-image) for the biomedical applications. We emphasize that our data generation strategy is highly scalable and opt to limit the dataset size to current scale due to budget constraints. In addition, we focus on equipping the mixed-modal model with additional skills to enhance the diversity of the MEDMAX dataset.

4.2 Dataset Curation for Diverse Skills

We aim to collect a diverse set of biomedical datasets to teach novel skills to the mixed-modal foundation models. We curate high-quality datasets for individual skills as follows:

Visual question answering (VQA): The ability to answer questions for the images is crucial for biomedical assistants. In this regard, we include biomedical visual questions that involve answering close-ended questions (yes/no), open-ended questions, and multiple-choice questions in MEDMAX. In this context, the biomedical image and question constitute the input (or context), while the ground-truth answer serves as the output (or response).

Prior work such as LLaVA-1.5 [38, 32] have included the data from the general-purpose VQA benchmarks in their instruction tuning mix to enhance the model's visual capabilities. Hence, we combine the training sets of popular VQA datasets including VQA-RAD [30], SLAKE [36], PathVQA [20], and PMC-VQA [73]. To embed the knowledge about different anatomical regions (20+ regions), we split the OmnimedVQA [22] dataset into a training (81K) and testing set (1K), and add the training set into the MEDMAX data. In addition to the diversity in the domains and question styles, these datasets are also a rich source of expert-annotated data (e.g., clinician-driven annotation of VQA-RAD) that is critical for building high-quality data. In total, we have **284K** VQA instances in the MEDMAX dataset.

Image captioning and generation: The ability to interpret and analyze the biomedical images is critical for accurate disease monitoring and diagnosis [46]. On the flip side, the capability to generate realistic

[1] We reserve the use of PMC-OA (train split) for image captioning and generation task, as discussed in §4.2.

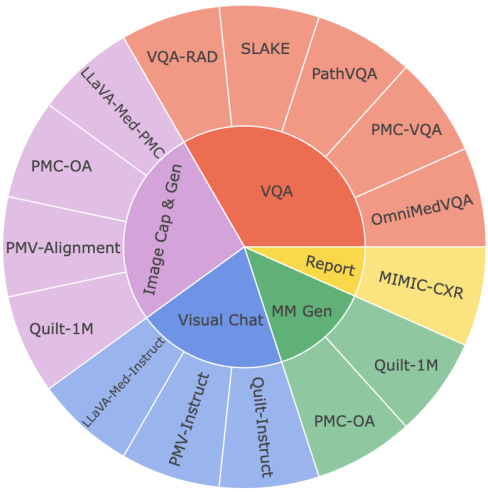


Figure 5: **Task-specific data sources in MedMax.** We present the data-sources used to curate task-specific data in the MedMax collection. We abbreviate PubMedVision as PMV, visual question answering as VQA, Image captioning and generation tasks as Image Cap & Gen.

biomedical images conditioned on the medical concepts, grounded in text, is useful for addressing the lack of high-quality annotated biomedical imaging data [13]. In this regard, a mixed-modal foundation model can benefit from potential knowledge transfer between both the tasks in the unified architecture. Here, we re-purpose a wide range of biomedical image-caption data found on the internet for the image captioning and generation task.

To this end, we collect image-caption data from four sources: (a) LLaVA-Med-PMC, (b) PMC-OA [35], (c) Quilt [24], and (d) PubMedVision-AlignmentVQA [71]. In particular, LLaVA-Med-PMC refers to the 600K subset of PMC-15M [72] (curated from PubMed Central research papers) that was released as a part of LLaVA-Med data. On manual inspection, we found that this data is mostly composed of statistical figures, graphs and plots on biomedical data instead of biomedical images (e.g., X-ray, MRI, CT). Since the focus of our work is to train models on biomedical data, we filter this data using BioMedCLIP score (Appendix C). Subsequently, we were left with 37K image-caption instances from LLaVA-Med-PMC data.

Further, we utilize the PMC-OA data that also consists of image-caption data curated from PubMedCentral research papers. Its notable features includes its accessibility (open-access), size ($1M+$ instances) and diversity of biomedical imaging data (e.g., ultrasound, fMRI, endoscope, PET). To maintain one-to-one correspondence in the data, we filter instances where a caption was aligned with multiple sub-images. Further, we filter small biomedical images that were less than 200 pixels in width or height in this data. Overall, the filtered training PMC-OA split contains 83K instances.

While the previous datasets are constructed from the biomedical research papers, we also include Quilt [24] in our dataset. Specifically, it consists 1M image-text pairs of histopathology dataset that is largely curated from the lectures on the youtube. We utilize 100K subset of this data for teaching for the captioning and generation subset of the MedMax data. Subsequently, we purpose these real image-text pairs for image captioning (image is the context and text is the response) and generation (text is the context and image is the response). We provide the finetuning templates in Appendix §D.

In many cases, the real image-caption has inherent data noise and formatting issues. As a result, we also include the synthetically-generated PubMedVision-AlignmentVQA data which utilizes GPT-4-Vision [44] to denoise and reformat internet data. In particular, we filter the original data (647K) to remove multiple image instances in this data to get 504K instances, and randomly select a subset of 100K instances for the MedMax mix. While this dataset can be utilized for image captioning (input image and question as context, and the answer as context), it could not be utilized for image generation directly. To address this, we prompt

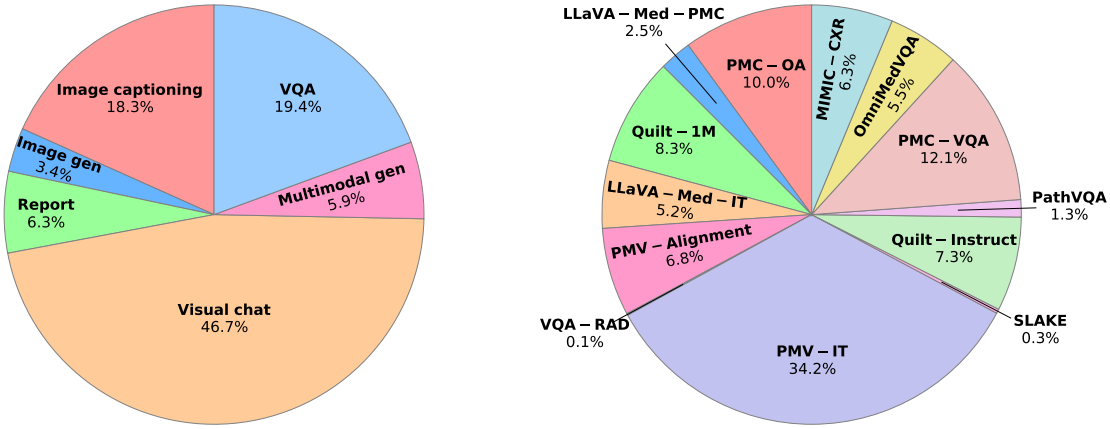


Figure 6: **Distribution of skills (left) and data sources (right) in MedMax.** We highlight that MedMax consists diverse skills for mixed-modal instruction tuning (left). In particular, we source the data from biomedical sources that cover several domains (e.g., radiology) and knowledge bases (e.g., research papers, YouTube).

GPT-4o-mini [23] to convert the image descriptions in the PubMedVision-AlignmentVQA data into image generation prompts: '*Convert the image description into an image generation prompt with AI*'. We show an example in Appendix Table 6. In total, we curate **320K** instances for the image captioning (160K) and the image generation (160K).

Visual chat: Biomedical visual instruction tuning [39, 33] enables foundation models to reason and solve diverse tasks related to biomedical images for the practitioners. Hence, we aim to curate a wide range of tasks that practitioners care about for various biomedical domains (e.g., radiology, histopathology) and sources (e.g., research papers and internet videos). Specifically, we utilize the existing data sources such as: (a) LLaVA-Med-Instruct [33], (b) Quilt-Instruct [49], and (c) PubMedVision-Instruct [71]. In particular, these datasets are created synthetically by prompting foundation models such as LLMs [5] with image descriptions or external knowledge (e.g., inline mentions in research papers or video transcripts).

Concretely, we collect 76K instances from LLaVA-Med-instruct-120K data. The remaining instances were not available publicly. In particular, half of this data was constructed using the image description and the other half also utilized the inline mentions of the images in the PubMedCentral research papers. To further enrich our data with diverse instructions, we include PubMedVision-IT dataset which is created using GPT-4-Vision [4]) as the data generator. Originally, this data consists 647K instances, but we filter instances with multiple images in the context that left us with 504K instances. Finally, we include 107K conversations from the Quilt-Instruct [49] to diversify our dataset with knowledge from the youtube videos (and transcriptions) for the histopathology domain. Overall, the MedMax consists **686K** instances for visual chat scenarios.

Medical report understanding: The ability to perform detailed inspection of a patient's imaging data requires specialized training. Thus, it is vital to expose our model to the expert-written findings (normal and abnormal anatomical cues in the image) from the patient's data. Hence, we collect chest radiographs along with the medical reports in the MIMIC-CXR [26] data. Originally, the dataset consisted of 377K instances. We filtered it to exclude reports discussing more than one image, reducing the dataset to 102K instances. Subsequently, we subsampled the data to decrease the proportion of 'No findings' reports [13] from 20% to 10%. Finally, we have **92K** instances consisting chest radiograph-report pairs. We purpose half of the dataset for radiograph-conditioned report generation task, and the other half to generate chest radiographs conditioned on the medical report. We provide the templates used for these tasks in Appendix D (Table 10

Table 1: **Additional information about diverse biomedical dataset sources.** We highlight that MEDMAX consists data across several biomedical domains and knowledge bases.

Data source	Domain	Knowledge Base
LLaVA-Med-PMC	Diverse	PubMed Central [2]
PMC-OA	Diverse	PubMed Central [2]
Quilt-1M	Histopathology	YouTube [3]
LLaVA-Med-IT	Diverse	PubMed Central [2]
PubMedVision-Alignment	Diverse	PubMed Central [2]
PubMedVision-IT	Diverse	PubMed Central [2]
Quilt-Instruct	Histopathology	YouTube [3]
VQA-RAD	Radiology	MedPix [1]
SLAKE	Radiology	MSD [7], CXR-8 [57], Chaos [28]
PathVQA	Pathology	PEIR Digital Library [27]
PMC-VQA	Radiology	PubMed Central [2]
OmniMedVQA	Diverse	Diverse
MIMIC-CXR	Chest X-ray	MIMIC-CXR [26]

and 11).

We present the details of the task-specific data sources in Figure 5. Further, we highlight the biomedical domains and knowledge bases covered in the MEDMAX in Table 1. Additionally, Figure 6 provides details on the biomedical domain, database, and the proportion of individual dataset sources. Furthermore, we note that MEDMAX comprises 725K unique images and 947K unique words. Collectively, these features underscore the diversity across various quality axes, making it well-suited for instruction tuning of mixed-modal foundation models.

5 Experimental Setup

Post data creation, we intend to instruction-tune a mixed-modal foundation model on the MEDMAX data (§5.1). Subsequently, we will present the evaluation methods to assess the performance of our model across diverse skills (§5.2).

5.1 MEDMAX Mixed-Modal Model

In this work, we will instruction-tune an instantiation of the Chameleon-7B [54] mixed-modal foundation model. Specifically, it can understand and generate multimodal content including images and text. Primarily, it represents the raw images as discrete visual tokens using a VQGAN [17], and the text data into discrete text tokens using BPE tokenizer [48]. Subsequently, each instance in the training dataset is represented as a sequence of discrete tokens, and the model is trained to predict the next token based on the preceding tokens in the sequence, following an autoregressive objective. The model consists of a vocabulary size of 65536 where 8192 are visual tokens apart from beginning of image and end of image tokens. In addition, the vocabulary includes a reserved token '<reserved08706>' that separates the instruction (context) from the response (output) for instruction-tuning. Post-tokenization, the entire MEDMAX data consists 1.7B tokens where 0.7B and 1B are visual and text tokens, respectively.

While the Chameleon-7B model weights are publicly available, they were safety-tuned and support mixed-modal inputs and text-only output to be used for research purposes.[2] To unlock the mixed-modal output capabilities, Anole [15] selectively finetunes the output embeddings of the image tokens using high-quality images from LAION [47]. This strategy does not interfere with the input mixed-modal understanding and text-only output abilities of the original Chameleon model. Hence, we choose Anole-7B as our base model for

[2] https://ai.meta.com/blog/meta-fair-research-new-releases/

Table 2: **Lists of tasks in the MEDMAX evaluation suite.** We perform comprehensive evaluation of the MEDMAX-finetuned mixed-modal model across diverse biomedical multimodal tasks. We note that BioMedCLIP can be used to assess the similarity between two images, which is referred to as Image-Image BioMedCLIPScore. We abbreviate visual question answering as VQA, multi-choice questions as MCQ, exact matching as EM, and large language model as LLM.

Task	Source	Metric
Biomedical Visual Question Answering		
VQA (Closed)	VQA-RAD [30]	Accuracy (EM)
VQA (Closed)	SLAKE [36]	Accuracy (EM)
VQA (Closed)	PathVQA [20]	Accuracy (EM)
VQA (Closed)	Quilt-VQA [49]	Accuracy (EM)
VQA (Open)	VQA-RAD [30]	Accuracy (LLM)
VQA (Open)	SLAKE [36]	Accuracy (LLM)
VQA (Open)	PathVQA [20]	Accuracy (LLM)
VQA (Open)	Quilt-VQA [49]	Accuracy (LLM)
VQA (MCQ)	PMC-VQA [73]	Accuracy (EM)
VQA (MCQ)	OmniMedVQA [22]	Accuracy (EM)
VQA (MCQ)	PathMMU [52]	Accuracy (EM)
VQA (MCQ)	ProbMed [65]	Accuracy (EM)
Biomedical Image Captioning and Generation		
Image captioning	PMC-OA [35]	BioMedCLIPScore
Image generation	PMC-OA [35]	BioMedCLIPScore
Image captioning	Quilt[24]	BioMedCLIPScore
Image generation	Quilt [24]	BioMedCLIPScore
Image captioning	MIMIC-CXR [26]	BioMedCLIPScore
Image generation	MIMIC-CXR [26]	BioMedCLIPScore
Biomedical Visual Chatbot	LLaVA-Med [33]	LLM score
Biomedical Multimodal Generation (NEW)	PMC-OA[35] Quilt[24]	LLM score Image-Image BioMedCLIPScore

mixed-modal instruction tuning in this work. In particular, we utilize low rank adaptation (LoRA) [21] for parameter-efficient finetuning of the base model. Further, we finetune the base model for 3 epochs on the MEDMAX dataset. We provide more finetuning details in Appendix G.

While various mixed-modal models, such as Transfusion [76] and Monoformer [75], offer different approaches, we select Chameleon (or Anole) for its simplicity in utilizing a single autoregressive loss, avoiding the complexity of balancing multiple objectives like autoregressive and diffusion. However, we note that MEDMAX can enable finetuning of any of these mixed-modal models too.

5.2 Evaluation

Post-training, it is critical to evaluate the capabilities of the instruction-tuned mixed-modal model across diverse skills. We provide the summary of the tasks and datasets in our comprehensive evaluation suite in Table 2. We will present the individual components here:

Biomedical VQA: This task involves answering complex questions about the biomedical images from different domains and anatomical regions of the human body. Specifically, we include the test set of VQA-RAD (radiology), SLAKE (semantic knowledge over radiology), PathVQA (pathology), and the entire QuiltVQA (histopathology) dataset. These datasets ask closed-ended (yes/no) and open-ended questions that require one word, phrase or sentence answer. Here, we use exact match to assess the accuracy of the models on the

closed-ended questions. However, the evaluation on the open-ended questions is intrinsically harder due to the subjectivity of the answers. Prior works [33] utilize the ratio of the ground-truth tokens appear in the generated answer, but this is quite unreliable when there are no constraints on the response format [37]. For instance, if the ground-truth is 'bronchi' and predicted answer is 'lung', then the metric will penalize the predicted answer. To avoid such challenges, we utilize an LLM (GPT-4o-mini) that compares the predicted answer against the ground-truth answer to decide where the model outputs are reliable or not. For each open-ended question, it gives a score of 0 or 1. We provide the evaluation template in Appendix Table 12.

Additionally, we also include medical VQA with multiple-choice questions datasets such as test set of the PMC-VQA (diverse biomedical domains), validation set of PathMMU (pathology) [52], and ProbMed (radiology) [65] dataset. In addition, we assess the performance of an hidden split of 1000 questions from the OmniMedVQA [22] dataset. Overall, we perform evaluations on **12** VQA tasks across diverse biomedical domains, skills, and question formats.

Biomedical image captioning and generation: Here, we aim to compare the capability of the MedMax model and the base Chameleon (Anole) modle to caption as well as generate biomedical images from different domains. In total, we collect 1200 instances from the testing split of PMC-OA (400), MIMIC-CXR (400), and unseen split of Quilt-1M (400) datasets. In particular, half of the dataset will be used for captioning evaluation and the other half will be used for generation evaluation. Similar to [13], we extract the impressions (summary of the findings) from the report data and treat them as the ground-truth captions for the associated chest radiographs in the MIMIC-CXR dataset. In principle, this will highlight the medical report understanding and report-conditioned image generation capability of our model. Subsequently, we will compute the BioMedCLIPScore [72] to assess the closeness between the input image and predicted caption (and vice-versa). We present the details for model inference in Appendix E.2.

Biomedical Visual Chatbot: LLaVA-Med introduced a visual chatbot evaluation where the task of the model is to answer 193 novel questions about 50 unseen biomedical images. Specifically, the questions belong to two category: conversation and detailed description of the images. Subsequently, an LLM scores the predicted answer and the GPT-4 written reference answer out of 10 conditioned on the question, image caption and additional image context. Finally, we compute the average relative prediction score as the ratio of the score for predicted answer and score for the reference answer. In our work, we leverage GPT-4o-mini for the scoring the generated responses due to its cost-effectiveness and better capabilities than GPT-4.[3]

Biomedical Multimodal Generation: With the development of mixed-modal biomedical assistants, it will be critical to assess their performance quantitatively. To this, we reserve 500 instances of the MedMax-Instruct data for testing and are never exposed during the training process. We note that each instance of this dataset consists of text input and multimodal (text and image) output. For a given text query, we prompt the base (or finetuned) model to generate interleaved response. Subsequently, we compare the text content in the predicted multimodal response with the reference text response using LLM-scoring, identical to the biomedical visual chat evaluation. In addition, we compare the generated image with the reference image using the image-image similarity score from the BioMedCLIP model. We present the evaluation templates and inference details in Appendix E.2.

We performed a data contamination analysis to verify the independence between our MedMax dataset and the test sets used in all downstream tasks. Our examination revealed no exact matches between image-text pairs across these datasets, confirming the absence of data leakage.

6 Experiments

In this section, we will present the downstream performance of the instruction-tuned mixed-modal model with the MedMax data. In particular, we provide the results for VQA tasks, image captioning and generation, multimodal generation, and visual chat (§6.1). To study the dataset scaling, we will analyze the performance by finetuning the mixed-modal on different subsets of MedMax (§6.2). Finally, we conduct ablation studies

[3] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

Table 3: **Performance of the MedMax model and baselines on the downstream VQA tasks.** We find that the MedMax mixed-modal model outperforms closed as well as open multimodal models on twelve VQA datasets. This highlights that the model can generalize well to unseen instances and tasks ranging across several biomedical domains.

Model	Chameleon (7B)	LLaVA-Med (v1.5-7B)	GPT-4o (mini)	GPT-4o	HuatuoGPT (Vision-7B)	MedMax (7B)
Average	39.4	36.6	38.5	42.0	52.4	65.5
VQA-RAD (Closed) [30]	48.6	61.0	55.8	54.2	74.5	75.3
SLAKE (Closed) [36]	59.1	48.7	50.4	50.1	70.7	88.4
PathVQA (Closed) [20]	58.9	62.7	48.7	59.2	65.9	91.8
QuiltVQA (Closed) [49]	71.4	63.0	38.5	44.6	55.7	61.2
VQA-RAD (Open) [30]	32.0	23.0	13.0	17.6	19.0	46.5
SLAKE (Open) [36]	5.3	25.1	49.3	63.7	53.3	82.2
PathVQA (Open) [20]	18.0	6.2	7.3	9.1	6.0	40.6
QuiltVQA (Open)[49]	15.3	17.2	28.0	36.1	22.2	26.0
PMC-VQA [73]	31.0	18.9	39.6	40.8	51.6	49
OmniMedVQA [22]	45.7	28.7	45.1	40.9	75.6	99.5
PathMMU [52]	34.5	29.8	35.6	39.1	55.4	49.3
ProbMed[65]	52.8	58.5	50.6	48.3	78.7	75.8

to understand the impact of the specialized visual encoder, and role of specific data subsets on downstream performance (§6.3).

6.1 Main results

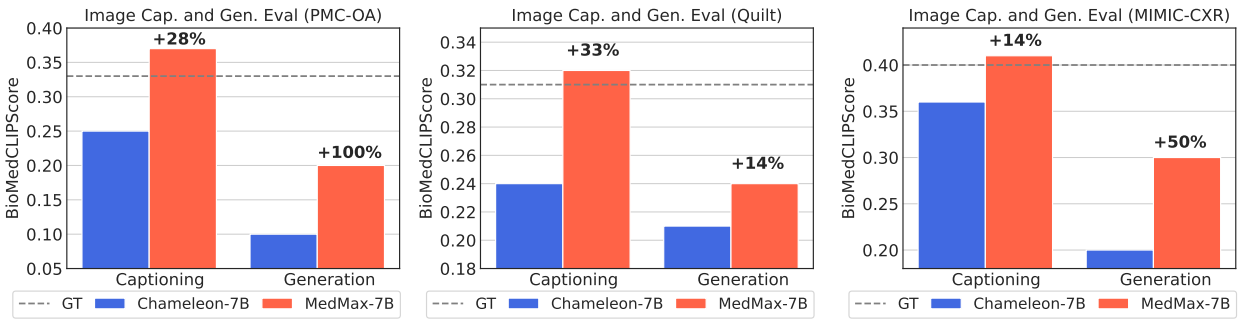
Here, we will present our empirical findings on the diverse downstream tasks.

6.1.1 Biomedical Visual Question Answering

We compare the performance of the finetuned MedMax model against several vision-language models on the battery of the VQA datasets in our evaluation suite. Specifically, the baselines includes the open models: the mixed-modal Chameleon-7B [54], and multimodal input but text-only biomedical specific models: LLaVA Med-v1.5 [39], and HuatuoGPT-Vision-7B [71]. In addition, we consider two closed vision-language models: GPT-4o-mini and GPT-4o. We present our results in the Table 3.

Our experiments reveal that MedMax model outperforms the base model (Chameleon) by 26.1% percentage points on the average VQA performance across twelve tasks. This highlights that MedMax is suitable for building biomedical specialized model from the base model via instruction-tuning. Moreover, we observe that it achieves the *best* average performance across diverse baselines. In particular, it beats the closed vision model, GPT-4o, by 18.3% percentage points. It indicates that the MedMax model is the most capable multimodal model for biomedical VQA across open as well as closed models. Further, we highlight the task-specific performance on individual VQA tasks too.

In addition, we observe that MedMax achieves the highest accuracy amongst the baselines for 7 out of 12 tasks. Interestingly, we find that Chameleon-7B achieves the highest QuiltVQA (Closed) but a low QuiltVQA (Open) accuracy which suggests that closed-ended questions are easier to solve for the base model. In addition, we observe that MedMax achieves a performance of 99.5% on the unseen split of OmniMedVQA, highlighting that this dataset is not particularly challenging. The model demonstrates the ability to answer questions about diverse anatomical regions after being exposed to some in-distribution data from OmniMedVQA. Furthermore, we note that HuatuoGPT-Vision-7B outperforms other baselines on 3 out of the 12 datasets. This can be attributed to its high-quality fine-tuning dataset and the Qwen2 backbone [66], which was pretrained on 7 trillion text tokens. This backbone significantly enhances the model's ability to comprehend diverse biomedical domains and formatting styles more effectively.



(a) Performance on PMC-OA. (b) Performance on Quilt. (c) Performance on CXR.

Figure 7: **Performance on the image captioning and image generation tasks.** We find that MedMax model consistency outperforms the base Chameleon mixed-modal model across diverse biomedical domains.

Additionally, we observe that the MedMax model achieves performance improvements of 10.7%, 15%, and 23% on the unseen QuiltVQA (Open), PathMMU, and ProbMed datasets, respectively. This underscores the MedMax model's ability to generalize effectively to novel downstream tasks through exposure to expert knowledge, question formats, and biomedical domains via instruction-tuning. Furthermore, we find that the MedMax is competitive to the task-specific finetuning of the LLaVA med on VQA-RAD, SLAKE and PathVQA datasets (Appendix F). This indicates that the practitioners can utilize an unified MedMax model instead of hosting separate task-specific LLaVA med models. Overall, these results highlight at the high-quality of the MedMax dataset for finetuning the mixed-modal foundation models for biomedical assistants.

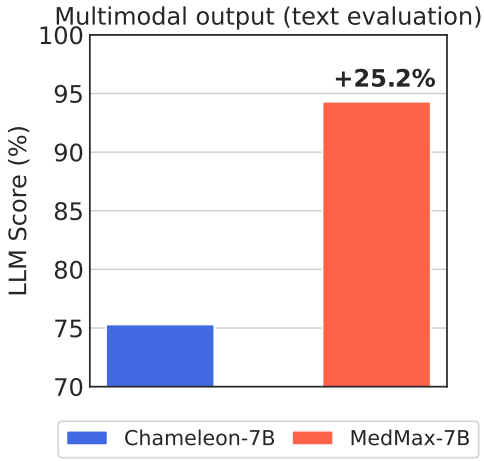
6.1.2 Biomedical image captioning and generation

Here, we compare the capability of the MedMax model and base model to interpret and generate biomedical image data across different biomedical domains. We present the results in Figure 7. Our empirical findings that the MedMax consistently outperforms the base model on the biomedical image captioning as well as the image generation task. In particular, the MedMax outperforms Chameleon by achieving a relative gain of 28%, 33% and 14% on biomedical captioning for the images in the PMC-OA, Quilt, and MIMIC-CXR datasets, respectively. Additionally, the MedMax also outperforms the Chameleon by achieving a relative gain of 100%, 14%, and 50% on the image generation tasks across PMC-OA, Quilt,and MIMIC-CXR captions, respectively.

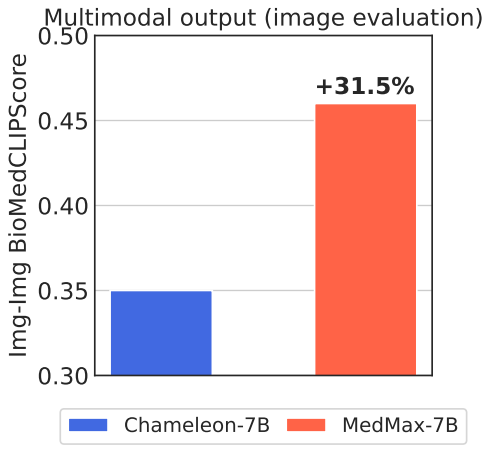
Further, we compare the captioning abilities of our model against other models in Appendix H. In particular, we observe that the average BioMedCLIPScore for our model is better LLaVA Med-v1.5 and at par with baselines such as HuatuoGPT-Vision, GPT-4o-mini, and GPT-4o. However, we highlight that none of the those models can perform image generation natively, unlike MedMax and Chameleon models. Overall, our results demonstrate that the MedMax model excels at reasoning about images and generating realistic biomedical visuals aligned with user intent.

6.1.3 Biomedical Multimodal Generation

Here, we aim to study the capability of the MedMax model to generate multimodal (interleaved image-text) content. We present the results in Figure 8. Our empirical findings highlight that the MedMax outperforms Chameleon by achieving a relative improvement of 25.2% on the quality of the text content in the multimodal output, as measured by LLM score. Furthermore, we observe that the MedMax outperforms Chameleon by achieving a relative improvement of 31.5% on the synthesized biomedical image in the multimodal output, as measured by the image-image similarity score using BioMedCLIP. This indicates that instruction tuning with MedMax data equips a mixed-modal model with strong multimodal generation capabilities in the



(a) Assessment of text content in MM output.



(b) Assessment of image content in MM output.

Figure 8: **Performance on the multimodal generation task.** Comparison between the performance of the MedMax and Chameleon mixed-modal model on the multimodal generation task. (a) We use LLM score to assess the quality of the generated data against the reference answer. (b) We use the image-image similarity score using BioMedCLIPScore to [72] compare the generated image and reference image. We find that MedMax finetuning improves the multimodal content generation capabilities for the biomedical domain.

biomedical domain. Overall, our results lay the foundation for future exploration in unlocking and evaluating multimodal generation capabilities with mixed-modal foundation models via instruction-tuning.

6.1.4 Biomedical Visual Chat

Here, we study the ability of the MedMax model and other baselines including Chameleon, LLaVA-Med-v1, LLaVA-Med-v1.5, and HuatuoGPT-Vision to answer novel queries about the biomedical images in the PMC dataset. We present the results in Figure 9. We observe that the overall LLM score of MedMax model is higher than the base Chameleon model and LLaVA med-v1 by 34% and 10.2% percentage points. This indicates instruction-tuning enables strong visual chatting capability to the mixed-modal models. In addition, the MedMax model is quite competitive in comparison to the best-performing model, LLaVA Med-v1.5, with a difference of 2.1% percentage points.

The fine-grained analysis indicates that the LLaVA med-v1.5 gains most of the performance on the conversation split of the dataset, while HuatuoGPT-Vision is better on the description split of the dataset. Such improvements in visual chat performance for LLaVA med-v1.5 and HuatuoGPT-Vision can be attributed to the quality of their language backbones, namely Mistral-v0.2-Instruct [25] and Qwen-2 [66], respectively. These backbones enhance the models' ability to interpret human queries and generate high-quality textual outputs.We believe that future advancements in base mixed-modal models will further enhance performance on biomedical visual chatting using the MedMax dataset.

6.2 Data scaling

Now, we explore how the benefits of mixed-modal instruction tuning scale with the size of the dataset. Specifically, we finetune the Anole-7B (instantiation of Chameleon-7B) with three subsets of the MedMax including 25%, 50%, and 75% of the data. Subsequently, we evaluate the average performance on the twelve VQA tasks for these subsets and the entire dataset. We present the performance vs the dataset scale in Figure 10. Our empirical finding suggests that the downstream performance of the instruction-tuned mixed-modal

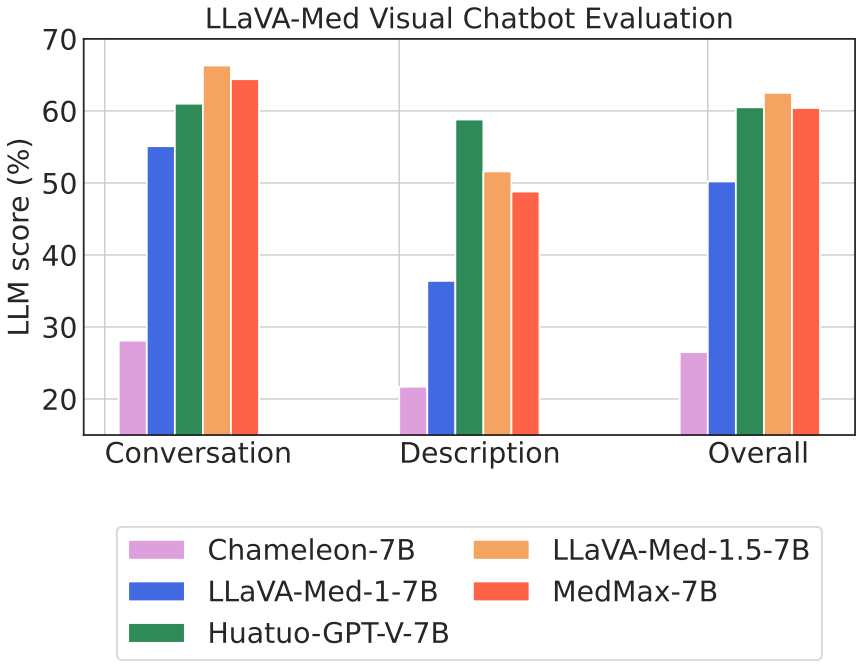


Figure 9: **Performance on the visual chat task.** Comparison between the visual chat performance of MedMax model and several baselines. We find that the chatting capabilities of our model is quite competitive, suggesting its ability to answer novel queries about biomedical images.

model monotonically increases with the size of the data. This highlights that the MedMax dataset is of high quality, and further scaling has the potential to yield greater improvements on downstream tasks.

6.3 Ablation Studies

In this section, our goal is to the study the role of different factors that can impact the downstream performance during mixed-modal instruction tuning.

6.3.1 Impact of Specific Data Subsets

Since MedMax is a collection of several tasks, we aim to understand the impact of specific tasks on downstream performance. To this end, we create two subsets of the MedMax dataset: (a) all tasks except VQA, and (b) all tasks except visual chat. We select VQA tasks because our model achieves very high performance on them, and visual chat tasks because they constitute the largest proportion of the dataset. Subsequently, we fine-tune the base model on these two ablated versions of MedMax. The results are presented in Figure 11.

In Figure 11a, we observe that the model fine-tuned on the VQA-ablated dataset suffers a performance drop of 23% compared to the original dataset on VQA tasks. This highlights the importance of including high-quality VQA tasks in the MedMax mixture. Similarly, Figure 11b shows that the model fine-tuned on the chat-ablated dataset experiences a performance drop of 17% on visual chat tasks. This underscores the critical role of visual chat data in achieving strong performance on visual chat tasks. Overall, our experiments demonstrate the value of incorporating diverse tasks in the data mixture to enable generalization across downstream biomedical tasks.

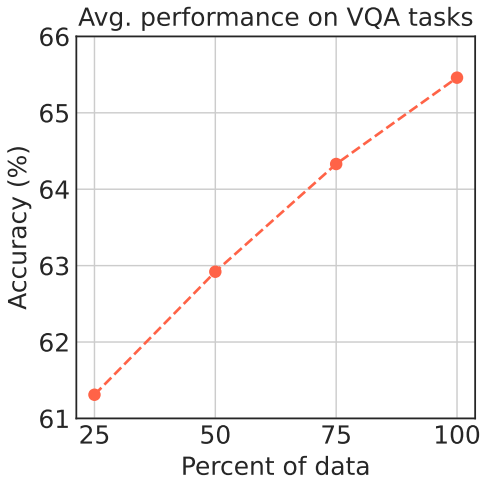


Figure 10: **Model performance with data scale.** We find that MedMax is a high-quality dataset that enhances VQA performance as it scales.

6.3.2 Impact of Specialized Visual Encoder

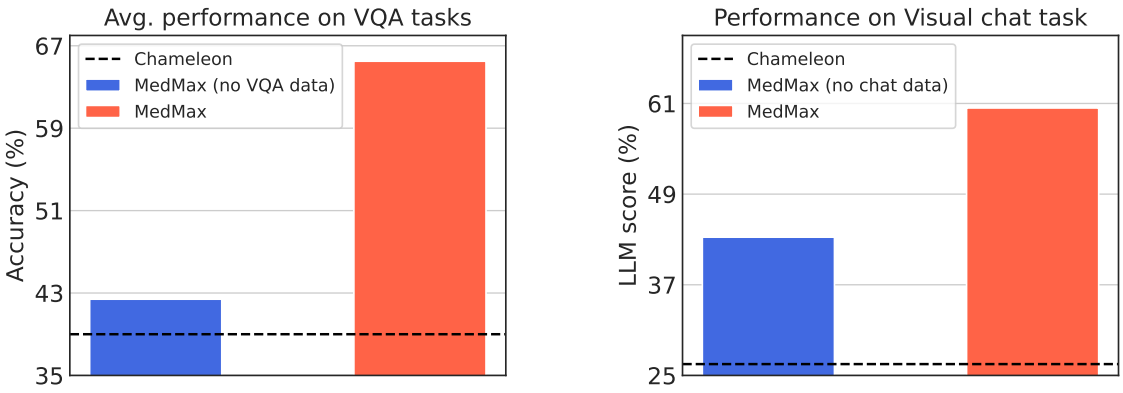
In the visual instruction tuning literature [33], a two-stage fine-tuning approach is commonly employed. In the first stage, the visual encoder is specialized to match the distribution of images in the instruction-tuning dataset, while keeping the transformer backbone of the architecture frozen. In the same spirit, we explore whether finetuning of the Chameleon's VQGAN encoder with biomedical images before instruction-tuning with MedMax leads to a better downstream model. In this regard, we finetune the base VQGAN with 300K images from the MedMax dataset for 8 epochs.[4] We find that the L1 reconstruction loss for the finetuned VQGAN was 7.8 in comparison to the base VQGAN 8.1 on a set of held-out biomedical images. This indicates that the fine-tuned encoder effectively represents the new domain better than the base visual encoder. Next, we select a random subset of 800K samples from the MedMax dataset and tokenize the images using the newly fine-tuned visual encoder. We then fine-tune the multimodal model using both the original subset and the re-tokenized subset under identical settings.

We present the performance of the two instruction-tuned mixed-modal models in Figure 12. We find that the model finetuned with the new discrete visual tokens achieves an inferior performance to the model finetuned with the original (base) visual tokens from the VQGAN. Specifically, we observe a gap of 3% averaged across the VQA datasets in our evaluation suite. This can be attributed to the distribution shift in the discrete visual tokens relative to the base model, where the base VQGAN visual tokens are better aligned with the base model's representations. Further exploration of fine-tuned, specialized visual encoders for discrete multimodal models is left for future work.

7 Related Work

Multimodal biomedical assistants. While early biomedical language models like ChatDoctor [34], MedicalGPT [63], and HuatuoGPT [71] advanced text-only medical reasoning (often built upon large language models such as LLaMA or Alpaca variants), they lacked multimodal capabilities for integrating visual information. This limitation prompted the development of multimodal biomedical models, ranging from encoder-only architectures like BiomedCLIP [72] to generative vision-language models capable of producing medical explanations. For instance, Med-Flamingo [42] extended OpenFlamingo [9] to a few-shot medical VQA paradigm via continued pre-training on curated image-text pairs. MedVInT [73], based on a pre-trained

[4]We used the original VQGAN codebase for finetuning: `https://github.com/CompVis/taming-transformers`.



(a) Data ablation for VQA task.

(b) Data ablation for visual chat task.

Figure 11: **Results for the data ablation study.** Finetuning the mixed-modal model with an ablated version of the MEDMAX data where the (a) VQA task instances and (b) visual chat instances are removed. The results highlight the usefulness of task-specific data in the mixture for downstream performance.

vision-language model, leveraged the PMC-VQA dataset to improve generative VQA. LLaVA-Med [33], built upon a LLaVA model [39], refined these capabilities using filtered PubMed data and GPT-4 generated instructions. RadFM [61] broadened image modalities to 2D and 3D radiology data. HuatuoGPT-Vision [14] adapts the LLaVA-v1.5 [38] architecture with Qwen2-7B [10] backbone and employs PubMedVision for large-scale medical VQA. Med-Gemini [55] integrated advanced multimodal and retrieval mechanisms on top of a Gemini model to enhance long-context and medical image understanding. More recently, MedTrinity-25M [62] proposed a benchmark of over 25 million image-ROI-description triplets that will be useful for pretraining. While MEDMAX has a smaller scale instruction-tuning data, it prioritizes efficient instruction-tuning through careful curation. While most existing approaches center on evaluation tasks like VQA or text-based medical chats, MEDMAX pushes beyond the boundary by demonstrating mixed-modal generation and interleaving text-image content to further enrich clinical comprehension.

Multimodal instruction tuning. Multimodal model training typically begins by aligning modalities in a shared embedding space and then perform instruction tuning to enhance conversational capabilities [12]. LLaVA [39] was among the first to utilize multimodal instruction-following data, generated by GPT-4, to enable rich visual conversations. MiniGPT-4 [77] constructed instruction sets by combining image-text datasets from Conceptual Caption [50], SBU [45], and LAION [47] with handwritten instruction templates, while InstructBLIP [16] incorporated VQA datasets to enhance visual reasoning. Multi-Instruct [64] further diversified the instruction set by incorporating 47 multimodal tasks. Beyond single images, MIMIC-IT [31], LAMM [67], and Macaw-LLM [41] introduced 3D, audio, and video scenarios for broader multimodal understanding. More recent datasets, such as LLaVAR [74], augmented visual instruction tuning with OCR results and expanded capabilities to handle text-rich images. High-quality instruction tuning data can be combined from multiple sources: LLaVA-1.5 [38] improved upon LLaVA [39] by incorporating diverse academic instruction tuning data, while LLaVA-OneVision [32] extended this approach by combining data across single-image, multi-image, and video scenarios. In this work, MEDMAX integrates multiple medical image datasets to create high-quality instruction tuning data that enables mixed-modal generation capabilities.

Mixed-modal foundation models. Mixed-Modal foundational models use a single neural network to process inputs of multiple modalities. The training objectives of such models comes in different flavors. Earliest works such as BEIT-3[56] make use of masked data modeling or contrastive learning objective from self-supervised learning field. More recent works uses generative modeling objectives instead. Among these, some work such as UniDiffuser [11] use a diffusion objective to learn a joint distribution of image and text

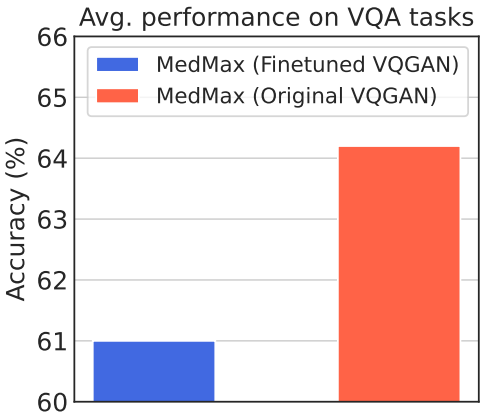


Figure 12: **Effect of finetuned visual encoder on downstream performance.** We find that the finetuning the Chameleon model with the original visual tokens gives better downstream performance than finetuned visual token on biomedical images. This highlights that the model does not like the distribution shift in the visual token distribution.

in latent space, and Transfusion [76] combines diffusion for images with autoregressive modeling for text. Alternatively, models such as Unified-IO [40], Chamaleon [54], Emu3 [58], CM3Leon [68], and Anole [15] formulate multi-modal learning as a general sequence modeling problem over multi-modal tokens. This autoregressive discrete decoding approaches facilitates generation of interleaved text-image sequences while maintaining architectural simplicity. In this work, we leverage this architectural simplicity of autoregressive mixed-modal models to effectively train MedMax, enabling comprehensive biomedical instruction tuning across diverse tasks and modalities.

8 Conclusion

In this work, we present MedMax, an instruction-tuning dataset designed for training mixed-modal foundation models. This dataset equips models with diverse multimodal biomedical capabilities across several domains. Notably, it is the first instruction-tuning dataset to enable multimodal generation capabilities for biomedical AI (MedMax-Instruct). Additionally, MedMax allows mixed-modal models to respond effectively to novel human instructions, functioning as capable biomedical assistants. Our experiments demonstrate that the MedMax-tuned model achieves competitive results across various downstream tasks, including biomedical visual question answering, multimodal generation, image captioning, image generation, and visual chatting. Overall, our work establishes a strong foundation for training the next generation of mixed-modal biomedical AI assistants.

Acknowledgements

AG would like to acknowledge an AI2050 Fellowship from Schmidt Sciences, NSF Career Award #2341040, and Amazon Research Award. HB is supported in part by AFOSR MURI grant FA9550-22-1-0380. SZ is supported in part by Amazon Fellowship. We would also like to thank Yidou Weng, Ethan Israel, Helen Cai, and Mohamed Soufi for their assistance in the qualitative assessment of our model.

References

[1] MedPix — medpix.nlm.nih.gov. `https://medpix.nlm.nih.gov/home`.

[2] PubMed Central (PMC) — pmc.ncbi.nlm.nih.gov. https://pmc.ncbi.nlm.nih.gov/.

[3] YouTube — youtube.com. https://www.youtube.com/.

[4] Gpt-4v(ision) system card. 2023.

[5] 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.

[6] Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784, 2022.

[7] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022.

[8] Aurora. Grok Image Generation Release — x.ai. https://x.ai/blog/grok-image-generation-release, 2024.

[9] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[10] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023.

[11] Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pages 1692–1717. PMLR, 2023.

[12] Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Marcella Cornia, Rita Cucchiara, et al. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*. 2024.

[13] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.

[14] Junying Chen, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, 2024.

[15] Ethan Chern, Jiadi Su, Yan Ma, and Pengfei Liu. Anole: An open, autoregressive, native large multimodal models for interleaved image-text generation. *arXiv preprint arXiv:2407.06135*, 2024.

[16] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[17] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021.

[18] Gemini-2.0 Flash. Gemini 2.0 Flash Experimental — deepmind.google. https://deepmind.google/technologies/gemini/flash/, 2024.

[19] Yu Gu, Jianwei Yang, Naoto Usuyama, Chunyuan Li, Sheng Zhang, Matthew P Lungren, Jianfeng Gao, and Hoifung Poon. Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. *arXiv preprint arXiv:2310.10765*, 2023.

[20] 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.

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[22] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.

[23] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[24] Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36, 2024.

[25] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[26] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

[27] Kristopher N Jones, Dwain E Woode, Kristina Panizzi, and Peter G Anderson. Peir digital library: Online resources and authoring system. In *Proceedings of the AMIA Symposium*, page 1075. American Medical Informatics Association, 2001.

[28] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

[29] Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. T\" ulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.

[30] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018.

[31] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning, 2023.

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

[33] 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. *Advances in Neural Information Processing Systems*, 36, 2024.

[34] Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6), 2023.

[35] 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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 525–536. Springer, 2023.

[36] 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. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021.

[37] Gang Liu, Jinlong He, Pengfei Li, Genrong He, Zhaolin Chen, and Shenjun Zhong. Pefomed: Parameter efficient fine-tuning of multimodal large language models for medical imaging, 2024.

[38] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[39] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[40] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action. *arXiv preprint arXiv:2312.17172*, 2023.

[41] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*, 2023.

[42] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-flamingo: A multimodal medical few-shot learner. July 2023. arXiv:2307.15189.

[43] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023.

[44] OpenAI. Gpt-4v(ision) system card, 2023b. `https://openai.com/research/gpt-4v-system-card`, 2023.

[45] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

[46] John Pavlopoulos, Vasiliki Kougia, and Ion Androutsopoulos. A survey on biomedical image captioning. In *Proceedings of the second workshop on shortcomings in vision and language*, pages 26–36, 2019.

[47] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[48] Rico Sennrich. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[49] Mehmet Saygin Seyfioglu, Wisdom O Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13183–13192, 2024.

[50] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[51] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

[52] Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, et al. Pathmmu: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *European Conference on Computer Vision*, pages 56–73. Springer, 2025.

[53] 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. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[54] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

[55] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[56] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

[57] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[58] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

[59] Zifeng Wang, Hanyin Wang, Benjamin Danek, Ying Li, Christina Mack, Hoifung Poon, Yajun Wang, Pranav Rajpurkar, and Jimeng Sun. A perspective for adapting generalist ai to specialized medical ai applications and their challenges. *arXiv preprint arXiv:2411.00024*, 2024.

[60] 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. *arXiv preprint arXiv:2109.01652*, 2021.

[61] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology, 2023.

[62] Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Zou, Cihang Xie, and Yuyin Zhou. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine, 2024.

[63] Ming Xu. Medicalgpt: Training medical gpt model. `https://github.com/shibing624/MedicalGPT`, 2023.

[64] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11445–11465, 2023.

[65] Qianqi Yan, Xuehai He, Xiang Yue, and Xin Eric Wang. Worse than random? an embarrassingly simple probing evaluation of large multimodal models in medical vqa. *arXiv preprint arXiv:2405.20421*, 2024.

[66] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[67] Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Lu Sheng, Lei Bai, Xiaoshui Huang, Zhiyong Wang, et al. Lamm: Language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. *arXiv preprint arXiv:2306.06687*, 2023.

[68] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, et al. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *arXiv preprint arXiv:2309.02591*, 2(3), 2023.

[69] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*, 2023.

[70] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*, 2024.

[71] Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*, 2023.

[72] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023.

[73] 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.

[74] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023.

[75] Chuyang Zhao, Yuxing Song, Wenhao Wang, Haocheng Feng, Errui Ding, Yifan Sun, Xinyan Xiao, and Jingdong Wang. Monoformer: One transformer for both diffusion and autoregression. *arXiv preprint arXiv:2409.16280*, 2024.

[76] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024.

[77] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Limitations

In this work, we focus on diverse multimodal biomedical skills, including VQA, multimodal generation, visual chat, image captioning, image generation, and report understanding. While these skills cover a broad spectrum of tasks, there remain additional possibilities that could further enhance biomedical applications. For instance, we do not address setups that involve understanding and generating multiple images, which are critical for applications such as counterfactual biomedical image generation [19] and reasoning from multiple images [69]. Achieving this capability presents significant challenges, including the lack of openly available, large-scale, high-quality multi-image biomedical datasets and the limited context length of the pretrained (base) Chameleon model. To bridge this gap, more efficient methods for representing image data are required, rather than always encoding all images as 1024 tokens, which occupy a substantial portion of the model's context length. We leave these explorations for future work.

Furthermore, we emphasize that the MedMax model is a research prototype designed to encourage community engagement in building capable biomedical assistants. In cases of medical emergencies, we strongly advise users to consult a healthcare professional rather than relying solely on the model's output. With ongoing community feedback and the collection of high-quality expert data, we aim to improve the model's faithfulness and safety over time.

B MedMax-Instruct Data Curation Prompts

We present the GPT prompt to filter the bad captions from the real image-caption data in Table 4. Further, we provide the prompt for generating multimodal generation conversation using GPT in Table 5.

C LLaVA-Med-PMC Curation

We curated a dataset of 37.8K medical images filtered from an initial pool of 538K images, sourced from the data released by LLaVA-Med [33], which originates from PMC-15M [72]. The initial dataset contained a significant number of statistical figures, as the images were extracted from research articles. To filter these out and retain only the desired medical images, we utilized the pretrained BiomedCLIP [72] model to classify images based on the taxonomy defined in PMC-15M.

Our focus was on retaining images from the classes "Magnetic Resonance," "CT," "X-Ray," "ECG," "Light Microscopy," "Dermatology," and "Endoscopy," which represent the primary topics used in LLaVA-Med. To determine class-specific confidence thresholds, we manually labeled 80 images and evaluated the model's predictive confidence using ROC curves, identifying optimal thresholds through Youden's J statistic to balance sensitivity and specificity. Additionally, we applied a heuristic to exclude images with high prediction scores for statistical figures within the top five predictions. This filtering process resulted in a high-quality dataset of 37.8K images, focused on key medical imaging modalities.

D Data creation templates for captioning, generation, and report understanding

Table 7 and Table 8 present complementary approaches to image captioning, with the former focusing on concise, brief descriptions and the latter encouraging comprehensive, detailed analyses of image content. Table 9 demonstrates various prompts for generating images from text descriptions, using diverse language

Table 4: **Prompt to assess the quality of the caption aligned with a biomedical image in the real image-caption data.**

Evaluate whether an image description provides substantive information by analyzing it against the following criteria:
1. Specificity: Does it contain precise details rather than vague descriptions?
2. Context: Does it provide relevant background or situational information?
3. Technical Details: Are any specific measurements, conditions, or technical terms included?
4. Purpose: Would the information be useful for professional analysis, decision-making, or documentation?

For medical descriptions specifically, consider:
- Anatomical details
- Condition characteristics
- Observable features
- Diagnostic relevance

Format your response as follows:
1. Analysis: Briefly explain why the description is or isn't informative (2-3 sentences)
2. Conclusion: End with either "The answer is: Yes" or "The answer is: No"

Example:

Description: Juvenile polyp or retention polyp is present.
Output: The description identifies a specific medical condition (juvenile/retention polyp) and confirms its presence, which is diagnostically relevant. While brief, this information is clinically useful for medical assessment and treatment planning.
The answer is: Yes

Note: Evaluate only the information provided in the description without making assumptions about missing details.

Description: [CAPTION]

to ensure accurate visual representation of textual input. Table 10 showcases prompts for medical report generation from diagnostic images, emphasizing structured radiological reporting formats and professional clinical observations. Table 11 illustrates prompts for generating medical images from clinical reports, focusing on accurate visualization of documented pathological findings and diagnostic features.

E Additional Evaluation Details

E.1 VQA Open-Ended Evaluation with LLM

Table 12 presents the template for evaluating the models on the open-ended questions of the VQA datasets. Our prompt is motivated from the GPT evaluation prompt in `https://github.com/jinlHe/PeFoMed/tree/main`.

E.2 Evaluation Templates and Generation Modes

We present the templates and generation modes for diverse tasks in our evaluation suite in Table 13. Following the approach used in Chameleon, we suppress the probability of visual tokens in the output to zero, ensuring that only text content is generated for VQA tasks. Additionally, 'image-gen' indicates that the probabilities for the text tokens are suppressed to zero to ensure that the model just generates an image in the response. Further, 'any-gen' highlights that the model is free to generate multimodal content in the response. We

perform greedy decoding in our experiments. Across our experiments, we use greedy decoding (temperature = 0) to generate text content and set the temperature to 0.7 for generating image content in the responses.

F Additional VQA Results

In Table 6.1, we compare various multimodal foundation models on VQA datasets. Our objective is to evaluate the performance of the MedMax model against the task-specific fine-tuning of LLaVA Med on diverse VQA datasets independently. The results for close-ended questions from the VQA-RAD, SLAKE, and PathVQA datasets are presented in Table 14.

We observe that MedMax outperforms LLaVA Med finetuned on the VQA datasets for three epochs, achieving improvements of 8.8% on VQA-RAD, 24.2% on SLAKE, and 2.3% on PathVQA. Furthermore, we note that MedMax performs better than LLaVA Med finetuned for 15 epochs on individual datasets for two out of the three VQA datasets (SLAKE and PathVQA).

These results highlight that a single MedMax model checkpoint is not only highly capable but also more practical for users, eliminating the need to maintain separate task-specific model checkpoints for popular VQA datasets.

G Finetuning Details

We fine-tune the Anole [15], an instantiation of Chameleon [54], on the MedMax dataset, which consists of 1.47 million instances. Specifically, we employ low-rank adaptation (LoRA) [21] for fine-tuning, using $r = 16$, $\alpha = 16$, and dropout = 0.05. The target modules include the {query, key, value, output, up, down, and gate} projection matrices. In total, this approach updates 40M parameters during fine-tuning. We train the model for 3 epochs using a cosine learning rate schedule (peak LR=1e-4 with a warmup ratio of 0.1) and a batch size of 8. The training is conducted on 8 Nvidia L40S GPUs (46GB GPU VRAM each).

H Additional Image Captioning Results

We compare the performance of the MedMax model with other relevant captioning models on the image captioning task using BioCLIPMedScore. The results are presented in Table 15. Our analysis reveals that the MedMax model outperforms LLaVA Med-v1.5 by an average of 0.16 points. Furthermore, we find that the performance of the MedMax model is on par with diverse baselines such as HuatuoGPT-Vision and GPT-4o for the captioning task. Overall, this underscores the capability of the MedMax model in training robust mixed-modal models that excel in diverse biomedical tasks.

Table 5: **Prompt to generate multimodal generation conversation using the caption in the real image-caption data.**

Task
Create natural, single-turn conversations that demonstrate how users might seek help understanding biomedical data descriptions without access to actual images.

Input Format
A brief, clinical description of biomedical data, focusing on: Measurements, Observed structures, Technical parameters, and Relevant findings.

Output Requirements

1. User Question
Generate a natural question that: Addresses specific medical findings or terminology, Avoids references to images, figures, or descriptions, Reflects a general understanding level, and Focuses on understanding clinical significance

Avoid phrases like: "In this image...", "Based on the description...", "According to the figure...", and "Can you explain what I'm seeing..."

Use formats like: Ẅhat does it mean when the common bile duct is 15mm?", "Can you explain what MRCP tells us about the pancreas?", "Is a dilated pancreatic duct concerning?"

2. AI Response
Structure the response with:

a) Clinical Interpretation: Begin with a clear, direct answer, Define medical terminology, Explain normal vs. abnormal values, Discuss clinical implications, Use accessible language while maintaining accuracy.

b) Visual Context: Insert '<image>' placeholder where relevant, Reference anatomical relationships, and Provide size comparisons to familiar objects when applicable.

c) Key Components: Diagnostic significance, Related conditions, Normal reference ranges, and Potential next steps or considerations

Style Guidelines

Language: Professional but accessible, Define technical terms, Use analogies when helpful, and Maintain clinical accuracy

Tone: Informative, Objective, and Reassuring without minimizing concerns

Structure: Clear topic sentences, Logical flow, Concise paragraphs, and Supporting details

Input: [CAPTION]

Table 6: **Converting the image description in the PubMedVision-AlignmentVQA to the image generation prompt using LLM.**

- **Image description in the PubMedVision-AlignmentVQA:** The image shows a chest radiograph in the anteroposterior (AP) view. The heart, mediastinal structures, and trachea appear to be displaced to the contralateral side, indicating dextrocardia, a condition where the heart is situated on the right side of the chest rather than the normal left side. The lung fields appear relatively clear, with no obvious abnormalities visible.
- **Image generation prompt using LLM:** Create an image of a chest radiograph in the anteroposterior (AP) view. Display the heart, mediastinal structures, and trachea displaced to the opposite side, illustrating the condition of dextrocardia, where the heart is located on the right side of the chest. Ensure the lung fields appear relatively clear and show no obvious abnormalities.

Table 7: **List of prompts examples for concise image captioning.**

- Describe the image concisely: [IMAGE]
- Provide a brief description of the given image: [IMAGE]
- Offer a succinct explanation of the picture presented: [IMAGE]
- Summarize the visual content of the image: [IMAGE]
- Give a short and clear explanation of the subsequent image: [IMAGE]
- Share a concise interpretation of the image provided: [IMAGE]
- Present a compact description of the photo's key features: [IMAGE]
- Relay a brief, clear account of the picture shown: [IMAGE]
- Render a clear and concise summary of the photo: [IMAGE]
- Write a terse but informative summary of the picture: [IMAGE]

Table 8: **List of prompts examples for detailed image captioning.**

- Describe the following image in detail: [IMAGE]
- Provide a detailed description of the given image: [IMAGE]
- Give an elaborate explanation of the image you see: [IMAGE]
- Share a comprehensive rundown of the presented image: [IMAGE]
- Offer a thorough analysis of the image: [IMAGE]
- Explain the various aspects of the image before you: [IMAGE]
- Clarify the contents of the displayed image with great detail: [IMAGE]
- Characterize the image using a well-detailed description: [IMAGE]
- Break down the elements of the image in a detailed manner: [IMAGE]
- Walk through the important details of the image: [IMAGE]

Table 9: **List of prompts examples for image generation from text descriptions.**

- Generate a visual representation based on the following description: [CAPTION]
- Create a depiction that accurately illustrates this description: [CAPTION]
- Generate an accurate representation aligned with this description: [CAPTION]
- Create a detailed depiction that reflects the information in this description: [CAPTION]
- Produce a clear visual based on the provided description: [CAPTION]
- Design a representation that captures the essence of the following text: [CAPTION]
- Generate a graphic aligned with this description: [CAPTION]
- Create an image that visualizes the details in the following text: [CAPTION]
- Develop a visual based on the description provided: [CAPTION]
- Illustrate the scenario described in the following text: [CAPTION]

Table 10: **List of prompts for image-to-report generation.**

- Generate a detailed medical report for this image following standard radiological reporting format.
- As a radiologist, provide a comprehensive medical report for this diagnostic image.
- Write a structured medical report describing all findings visible in this image.
- Examine this medical image and document your observations in a standard clinical report format.
- Create a detailed clinical report based on your analysis of this diagnostic image.
- Review this medical image and generate a complete radiological report including all relevant findings.
- Analyze this diagnostic image and provide a structured medical report with your observations.
- Acting as an experienced radiologist, document your interpretation of this image in a medical report.
- Evaluate this medical image and create a comprehensive clinical report detailing all findings.
- Provide a thorough radiological report based on your examination of this diagnostic image.

Table 11: **List of prompts for report-to-image generation.**

- Generate a medical image that accurately represents all findings described in this report.
- Create a diagnostic image that visualizes all the clinical observations mentioned in this report.
- Synthesize a medical image that corresponds to the findings detailed in this radiological report.
- Based on this clinical report, generate a medical image showing all described features and abnormalities.
- Produce a diagnostic image that illustrates all the medical findings documented in this report.
- Create a medical image that faithfully represents the pathological findings described in this report.
- Generate a diagnostic image that matches all the clinical observations in this medical report.
- Visualize this medical report as a diagnostic image showing all mentioned findings and characteristics.
- Transform this radiological report into a corresponding medical image with all described features.
- Based on the clinical descriptions in this report, generate an accurate medical image representation.

Table 12: **Template for evaluating the correctness of the predicted answer in comparison to the ground-truth answer for the open-ended questions in the VQA datasets.**

Given a question about an medical image, there is a correct answer to the question and an answer to be determined. If the answer to be determined matches the correct answer or is a good enough answer to the question, output 1; otherwise output 0. Evaluate the answer to be determined (1 or 0).

Question:
question about the medical image: **[question]**

Answers:
correct answer (ground truth): **[true answer]**
answer to be determined: **[generated answer]**

Task:
Given a question about an medical image, there is a correct answer to the question and an answer to be determined. If the answer to be determined matches the correct answer or is a good enough answer to the question, output 1; otherwise output 0. Evaluate the answer to be determined (1 or 0).

Output Format:
Correctness: **[your judgment]**

Table 13: **Template and generation modes for the downstream evaluation of the MedMax model.**

Task	Template	Generation Mode
VQA-RAD (Open/Closed) PathVQA (Open/Closed) SLAKE (Open/Closed) ProbMed	*<image>[question]*	Text
QuiltVQA (Closed)	*<image>Answer the question based on this image and respond 'yes' or 'no'.* *[question]*	Text
QuiltVQA (Open)	*<image>Answer the question based on this image.* *[question]*	Text
PMC-VQA OmniMedVQA PathMMU	*<image>[question]* *[choice A]* *[choice B]* *[choice C]* *[choice D]*	Text
Captioning	*<image>Please describe this picture.*	Text
Generation	*<caption>*	Image
Multimodal generation	*<question>*	Any
Visual chat (Conversation)	*<image>[question]*	Text
Visual chat (Description)	*<image>Analyze the image in a comprehensive and detailed manner.*	Text

Table 14: **Comparison between MedMax model and the task-specific finetuned LLaVA Med models on the closed-ended questions of VQA-RAD, SLAKE, and PathVQA datasets.**

	VQA-RAD (%)	SLAKE (%)	PathVQA (%)
LLaVA-Med-Finetuned (3 epochs) [33]	66.5	64.2	89.5
MedMax (Ours)	75.3 (+8.8)	**88.4** (+24.2)	**91.8** (+2.3)
LLaVA-Med-Finetuned (15 epochs) [33]	**84.2** (+17.7)	85.3 (+21.1)	91.2 (+1.7)

Table 15: **Comparison between MedMax model and other baselines on the biomedical image captioning using BioMedCLIPScore.**

	Average	**PMC**	**Quilt**	**MIMIC-CXR**
LLaVA-Med-v1.5 [33]	0.22	0.25	0.25	0.17
MedMax (Ours)	0.38	0.37	0.36	0.41
HuatuoGPT-Vision [71]	0.38	0.38	0.36	0.40
GPT-4o [23]	0.38	0.40	0.37	0.38