
ReMUD: A Reasoning Multimodal Ultrasound Dataset and Pipeline Brigding to General Specific Domains

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 Multimodal large language models (MLLMs) have shown great potential in general
2 domains but perform poorly in some specific domains due to a lack of domain-
3 specific image-text pairs. In some specific domains, there is abundant graphic
4 and textual data scattered around, but lacks standardized arrangement. Such as
5 in the field of medical ultrasound, there are ultrasonic diagnostic books, ultra-
6 sonic clinical guidelines, ultrasonic diagnostic reports, and so on. However, these
7 ultrasonic materials are often saved in the forms of PDF, images, etc., and cannot
8 be directly used for the training of MLLMs. This paper proposes a novel
9 image-text reasoning supervised fine-tuning data generation pipeline to create
10 specific domain quadruplets (image, question, thinking trace, and answer) from
11 domain-specific materials. A medical ultrasound domain dataset ReMUD is es-
12 tablished, containing over 45,000 reasoning and non-reasoning supervised fine-
13 tuning Question Answering (QA) and Visual Question Answering (VQA) data.
14 The Qwen2.5-VL-Ultra-7B model, fine-tuned on Qwen2.5-VL-7B-Instruct, out-
15 performs general-domain MLLMs in the medical ultrasound field. To facilitate
16 research, the ReMUD dataset, data generation codebase, and Qwen2.5-VL-Ultra-
17 7B model parameters will be released. The proposed approach provides a simple
18 and automated way to construct specific domain image-text data, addressing the
19 data shortage issue in specific domain MLLMs.

20 1 Introduction

21 In general domains, image-text data exists in abundance, like web images along with their corre-
22 sponding captions and contexts. Benefit from this, multimodal large language models (MLLMs)
23 instructions-tuned by leveraging multimodal inputs, such as LLaVA [27] and GPT-4o [31], have
24 displayed remarkable zero-shot task completion performance across a wide range of user-oriented
25 vision-language tasks, for instance, image understanding and reasoning. The rapid development of
26 MLLMs has provided more possibilities for expansion into broader fields and application scenar-
27 ios [38, 25, 7].

28 Despite demonstrating excellent vision-language abilities in general domains, MLLMs' perfor-
29 mance in specific domains often shows notable deficiencies [3]. Training data generated by publicly
30 available web data usually lacks domain-specific image-text pairs, resulting in a lack of domain-
31 specific expertise. General domain MLLMs may produce incorrect responses or complete hallucina-
32 tions in specific domain, such as medical field, which are not allowed to appear [5]. Pure text-based
33 models can be self-supervised through autoregression, predicting the next word. However, in con-
34 trast, MLLMs require the pre-preparation of relevant image-text pairs for training. As a consequence,

35 preparing image-text pair data for multimodal training demands substantial human and financial re-
36 sources, especially in some particularly specialized fields. Thus, we ask: Do we have a simple and
37 automated way to construct a large amount of image-text data from specific domain?

38 In this paper, we propose a novel image-text data generation pipeline to create specific domain
39 quadruplets (image, question, thinking trace and answer) from scratch. In a specific domain, there
40 are a large number of professional books and papers, which are often stored in the form of PDF.
41 These materials contain a wealth of curated image-text information. For example, the images in the
42 books, along with their corresponding text descriptions, are scattered throughout these resources.
43 Taking the medical ultrasound domain as an instance, we utilized the bounding box function of
44 Qwen2.5-VL [4] and the multimodal image-text recognition and generation capacities of GPT-4o
45 to establish a multimodal dataset **ReMUD** specific to the ultrasound domain. Subsequently, by
46 performing supervised fine-tuning with Qwen2.5-VL-7B-Instruct serving as the base model, the fine-
47 tuned model Qwen2.5-VL-Ultra-7B was developed. This model equipped with reasoning capability
48 demonstrates performance surpassing that of general-domain multimodal large language models
49 within the specialized realm of medical ultrasound. Specifically, our paper makes the following
50 contributions:

- 51 • **Specific domain quadruplets generation pipeline.** We present a novel data generation
52 pipeline to create domain specific quadruplets (image, question, thinking trace and answer)
53 from domain specific materials which can ensure the accuracy of professional knowledge.
54 This requires no manual annotations or pre-generated dataset and it is applicable to various
55 general specific domains.
- 56 • **ReMUD.** We establish the first open-source medical ultrasound domain dataset ReMUD,
57 which contains 45,000+ reasoning and non-reasoning supervised fine-tuning QA and VQA
58 data in the format of ShareGPT. It incorporates text Question Answering data, Visual Ques-
59 tion Answering data, reasoning data, and non-reasoning data. Three test datasets USQA-
60 Knowledge and UVQA-diagnos have been compiled to evaluate the ultrasound capabilities.
61 The reasoning model Qwen2.5-VL-Ultra-7B, which is fine-tuned on ReMUD, outperforms
62 general-domain MLLMs.
- 63 • **Open-source.** To facilitate research of specific domain and medical ultrasound field, we
64 will release the following assets to the public: the ReMUD instruction-following dataset,
65 the codebase for data generation and the parameters of Qwen2.5-VL-Ultra-7B.

66 2 Related work

67 **Multimodal large language model** Multimodal Large Language Models (MLLMs) have wit-
68 nessed a surge in research recently [40, 4, 42, 49, 39]. In architecture, they typically integrate
69 pre-trained modality encoders (e.g., ViT visual encoder and its variants [12, 34, 28]), pre-trained
70 LLMs (such as LLaMA [41], Vicuna [10], and Qwen series [33, 44]), and modality interfaces for
71 multimodal interaction [23]. Training MLLMs involves pre-training with large scale image-text
72 data for modality alignment [27], instruction-tuning to enhance task generalization, and alignment
73 tuning to match human preferences. Evaluation methods include closed-set evaluations on task-
74 specific datasets and open-set assessments like manual scoring, model scoring, and case studies.
75 Additionally, extended techniques such as Multimodal In-Context Learning (M-ICL) [13], Multi-
76 modal Chain of Thought (M-CoT) [48], and LLM-Aided Visual Reasoning have been developed
77 to enhance MLLMs’ capabilities. However, challenges remain, including handling long-context
78 multimodal information [47] and improving instruction-following [32], presenting opportunities for
79 future research. What’s more, LLaVA-Ultra [16] is a ultrasound multimodal large language model
80 that trained by professional ultrasound multi-modal data but not open-sourced.

81 **Visual Question Answering data generation** Manually creating Visual Question Answering
82 (VQA) datasets is a time-consuming and expensive process [35]. Template-based methods have
83 been developed to address VQA data shortage. They generate QA pairs using templates, creating
84 datasets like CLEVR [20] and MIMIC-Diff-VQA [18]. But models trained on such data perform
85 poorly on complex human-written questions [36]. Later, some studies use image descriptions and
86 LLMs to generate questions and answer by using LLMs’ high contextual understanding and reason-
87 ing ability to answer the question. For instance, Path-VQA [17] is obtained by extracting pathology

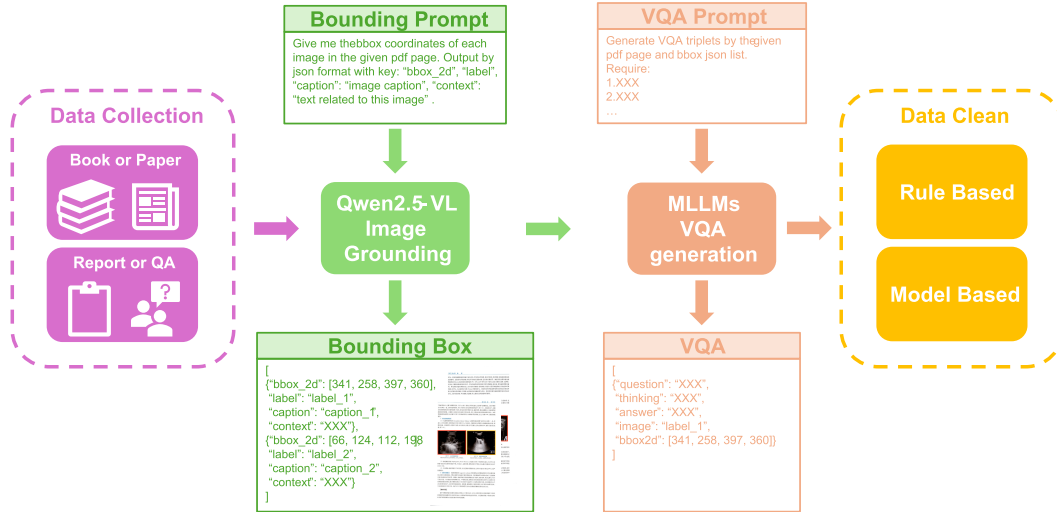


Figure 1: Flowchart of VQA Data Generation.

88 images and captions from publicly-available textbooks and digital libraries, using natural language
 89 processing techniques to generate question-answer pairs from the captions. Recently, there are even
 90 more MLLMs [31, 1] to assist in the generation of VQA data.

91 **Test-time reasoning** Test-time reasoning research has explored diverse methods to scale test-time
 92 compute. Parallel methods like majority voting [37, 6] and Best-of-N [21, 19] generate multiple so-
 93 lutions simultaneously, yet may not fully tap into a model’s reasoning. Sequential approaches allow
 94 for iterative refinement like Monte-Carlo Tree Search (MCTS) [15] and guided beam search [43].
 95 OpenAI’s o1 model [2] spurred interest in test-time scaling, with replication efforts like DeepSeek-
 96 R1 [11], though openly replicating its scaling behavior was challenging. Other research focused on
 97 improving reasoning via continued training on specialized corpora, new training methodologies, and
 98 prompting techniques [45, 46, 29].

99 3 Specific domain knowledge curation to create ReMUD

100 In this section, we introduce our process for collecting a large scale multimodal data in medical ul-
 101 trasound field and then creating **ReMUD** by automated image-text VQA generation. This is the first
 102 dataset for multimodal large language models specifically tailored to medical ultrasound. ReMUD
 103 is mainly in Chinese, with a small amount of English content.

104 3.1 Data generation from scratch

105 We found that there are few publicly available ultrasound data that involve ultrasound multimodality,
 106 i.e., datasets with ultrasound image-text pairings. Therefore, we constructed the ultrasound dataset
 107 from scratch, e.g., classic ultrasound textbooks, up-to-date ultrasound guidelines, open source ultra-
 108 sound dataset, and ultrasound paper, as a way to construct **ReMUD**. We generate **ReMUD** with two
 109 categories.

110 **Text-only data** Text-only data predominantly comprises textual ultrasound knowledge sourced
 111 from a diverse array of materials, including medical books, cutting-edge research papers, and in-
 112 depth technical documents. These materials serve as rich reservoirs of information, encompassing
 113 a wide spectrum of ultrasound concepts, ranging from fundamental principles to the latest advance-
 114 ments in the field. The data processing process is divided into three steps as follows:

- 115 1. First, we conducted an extensive crawl of publicly available data including medical
 116 books, cutting-edge research papers, and in-depth technical documents on the Internet to
 117 strengthen the model’s understanding of ultrasound basics.

Context:

Encapsulated pleural effusion: The pleural effusion is located between the chest wall and the lung, confined to a certain area, forming round, oval or semi-lunar anechoic areas of varying sizes. The base on the side close to the chest wall is relatively wide, and both ends are in an acute angle shape. The cavity wall thickens, and the inner wall is mostly uneven. Sometimes, there are septa within the cavity, and thick punctate or strip-shaped echoes can be seen. The liquid shows no signs of fluidity (Figure 7-12)

Bbox:

```
[{"bbox_2d": [341, 258, 397, 360], "label": "label_1", "caption": "caption_1"}, {"bbox_2d": [66, 124, 198, 240], "label": "label_2", "caption": "caption_2"}]
```



Response type 1: conversation

Question: What section is this ultrasound image taken from?

Answer: This is a coronal section image of a cranial ultrasound.

Label: label_1

Response type 2: complex reasoning

Question: Based on the ultrasound characteristics shown in this image, what is the most likely diagnostic result?

Thinking: In the image, the label "T" represents a tumor, and "R - PLE" represents right - sided pleural effusion. This indicates that there are two main ultrasound characteristics in the thoracic region, namely a space - occupying lesion (tumor) and accumulation of fluid in the thoracic cavity (pleural effusion). When a tumor appears in the thoracic cavity accompanied by pleural effusion, considering common diseases, situations such as tumor invasion of the pleura or the body's reaction induced by the tumor may lead to the development of pleural effusion. Among various lung and thoracic - related diseases, lung cancer is a relatively common disease that is prone to presenting in this way. Lung cancer cells can invade the pleura, increasing its permeability, thereby causing fluid to exude and accumulate, forming pleural effusion. Meanwhile, the tumor itself can also present as an abnormal space - occupying lesion in the thoracic cavity under ultrasound.

Answer: Overall, the hypoechoic mass with blood flow signals in the right thoracic cavity, along with ipsilateral pleural effusion, highly suggests the possibility of a lung tumor. Of course, the final diagnosis still requires integration with clinical findings and other examination results.

Label: label_2

Figure 2: An example for VQA generation.

2. Second, for formatted data, such as curated ultrasound knowledge question banks and widely recognized open-source ultrasound datasets, we adopted a structured conversion process. These data sources, despite already having a certain degree of organization, needed to be in a format that our model could easily process. We applied specific rules to transform them into the JSON format. On the other hand, there were also unformatted data that presented unique challenges. These data were stored in PDF or image formats, which are not directly amenable to text-based analysis. To overcome this hurdle, we turned to Optical Character Recognition (OCR) methods [14, 22]. OCR technology is capable of converting the visual text within PDFs and images into machine-readable text.
 3. After successfully converting the unformatted data into text format, we harnessed the power of the GPT-4o API. This advanced language model API was used to generate triplets consisting of a question, a thinking trace, and a generated answer. We set a strict requirement that these generated triplets must be derived entirely from the textual information provided. This ensures that the generated content is relevant and reliable.
- The questions and answers can take various forms, such as multiple-choice questions, which are useful for testing the model's knowledge recall and decision-making ability, or dialogues, which simulate real-world interactions and can better evaluate the model's understanding and in the ultrasound domain.
- Image-text data** In the process of creating multimodal datasets, the acquisition and processing of image-text data present a significant challenge. This is because generating high-quality image-text data usually demands a substantial amount of manual labor. Each image needs to be meticulously analyzed, and relevant text information, such as descriptions, captions, and annotations, must be accurately associated with it. This manual labeling work is not only time-consuming but also requires

141 a high level of expertise. For instance, in a medical multimodal dataset, where images might be
142 X-rays or MRIs, medical professionals are often needed to provide accurate text descriptions, which
143 further complicates the process and increases the cost. Here we propose an automated annotation
144 approach for the generation of image-text VQA data (See Fig 1):

- 145 1. Firstly, we use the precise object grounding function of Qwen2.5-VL to generate bounding
146 boxes of images in each page of a PDF. Each bounding box corresponds to a piece of JSON
147 data containing bounding box coordinates, labels and image caption.
- 148 2. Second, we take the pages with the generated bounding box labels and their corresponding
149 JSON data and pass them to powerful multimodal Large Language Model APIs, such as
150 GPT-4o or Gemini-2.0-Flash-Thinking-Exp [1]. Through the use of carefully crafted spec-
151 ific prompts, these APIs are able to generate Visual Question Answering (VQA) triplets
152 related to each bounding box. These triplets consist of a question about the image content
153 within the bounding box, a corresponding answer, and a thinking trace.
- 154 3. Finally, to complete the establishment of a comprehensive image-text dataset, we utilize
155 the bounding boxes to save the corresponding images. This step is essential for realizing
156 the one-to-one correspondence between the image and the text data. By ensuring this cor-
157 respondence, we can create a well-organized multimodal dataset where each image has its
158 associated text information, including the bounding box details, labels, captions, and VQA
159 triplets.

160 From above, we get quadruplets each containing a image path, a question, a thinking trace and an
161 answer.

162 Similar to DeepSeek-R1 [11], we divide the dataset into reasoning-data and non-reasoning data.
163 Specifically, for the generated factual QA dialogues, we do not generate the thinking trace, and for
164 the generated multiple-choice questions or diagnostic type questions, we generate the thinking trace
165 conform to the provided data.

166 3.2 Data cleansing and supplementation

167 We produced a large number of text-only and image-text data through automated methods, but these
168 data may contain errors and incompleteness, so we cleaned and supplemented the data according to
169 the following three principles: Quality, Specialty, and Diversity.

170 **Quality** To ensure data quality, we first remove incorrect data from API response errors and data
171 with non-existent image addresses. Subsequently, we employ a pre-trained binary classification
172 network to assess the legitimacy of images saved via bounding boxes. Finally, we utilize the Gemini-
173 2.0-Flash-Thinking-Exp API to evaluate if the generated triplets are supported by page information
174 and if the image-text match is of high quality through a scoring system, removing low-quality data
175 to enhance the dataset’s reliability for model training and evaluation.

176 **Professionalism** In line with the expert advice of professional doctors, we meticulously curate
177 professional ultrasound data by leveraging a diverse array of authoritative resources, including spe-
178 cialized professional books, cutting-edge research papers, and comprehensive clinical guidelines.
179 When making API calls, we incorporate specific requirements within the prompt words. Specifi-
180 cally, we stipulate that the generated answers must be traceable back to the data we have provided.
181 This ensures the accuracy and relevance of the output, maintaining a high standard of data qual-
182 ity. Moreover, when dealing with publicly accessible ultrasound multiple-choice question bank data
183 that features only one correct answer, we adopt a multi-step verification process. First, we utilize the
184 capabilities of GPT-4o to obtain its proposed answers and detailed explanations for each question.
185 After that, we carefully sift through these responses, retaining only those answers and explanations
186 that are correct. To further validate the retained content, we subject them to a rigorous review by
187 professional doctors. This combination of AI-assisted analysis and expert human judgment helps us
188 to create a reliable and high-quality ultrasound dataset for subsequent research and application.

189 **Diversity** We collect data encompassing diverse ultrasound domains like breast, thyroid, and fetal
190 ultrasound. Additionally, we gather publicly accessible ultrasound datasets from the Internet, such as
191 those for breast nodule benign and malignant classifications and ultrasound report generation [24, 9],

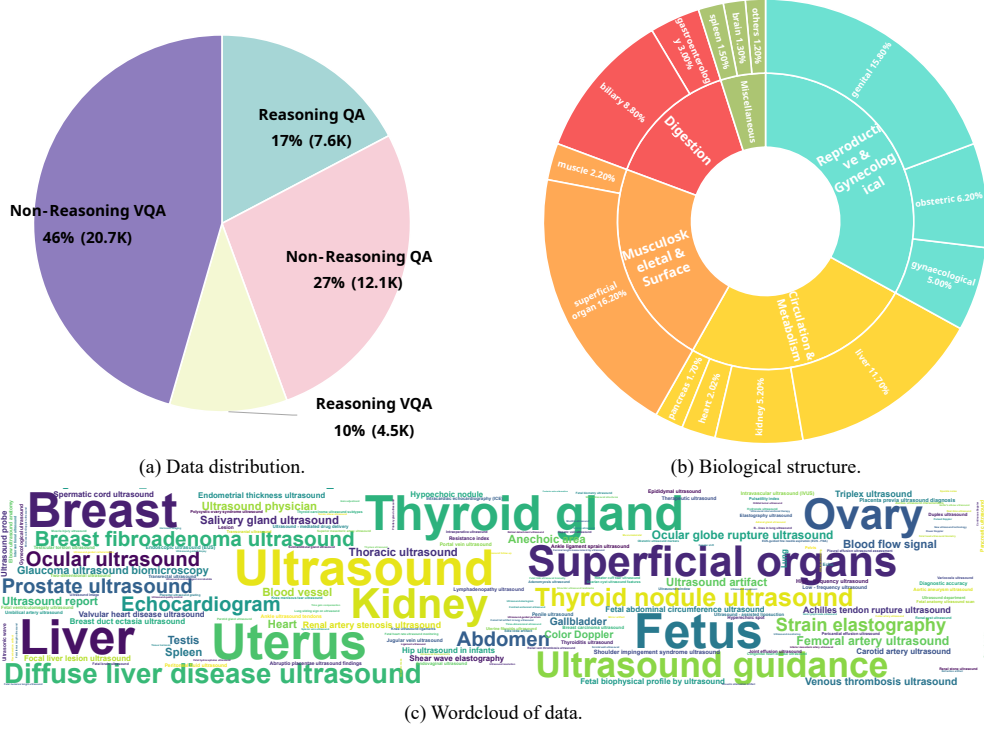


Figure 3: Statistical overview of ReMUD.

and transform them into triplets for better utilization. Moreover, we filter relevant information from publicly available medical QA datasets, Path-VQA [17] and PMC-OA [26], based on ultrasound keywords and incorporate it into **ReMUD**, aiming to create a comprehensive and high-quality dataset for in-depth research and application development in the ultrasound field.

3.3 Final selection of ReMUD

Through above methods, we get **ReMUD** containing text-only and image-text triplets. In order to evaluate the model’s capability in ultrasound field, We have created three test sets based on different data sources: USTQ-Open, USQA-Knowledge and UVQA-Diagnosis. USTQ-Open is compiled from publicly available Chinese ultrasound medicine question banks on the Internet. USQA-Knowledge automatically generated from ultrasound books. UVQA-Diagnosis is compiled based on the publicly available dataset [24]. Given the potential presence of redundant data, we utilize n-gram (n=12) analysis and a sentence transformer to efficiently identify and eliminate duplicates. This step is crucial as it ensures the uniqueness and quality of the test data, which in turn leads to more reliable evaluation results. These test data provide a broader range of scenarios and challenges for the model to encounter. Ultimately, **ReMUD** contains 45k+ training data and 1k+ test data with totally 20k+ images. The test data are in the form of multiple-choice questions and are accompanied by thinking trace. This structure allows us to not only evaluate the model’s ability to select the correct answer but also understand the reasoning process it undertakes, thereby facilitating a more in-depth and comprehensive assessment of the model’s capabilities in the ultrasound field.

4 Method

We used the Qwen2.5-VL-7B-Instruct model [4] as our base model. On this basis, we performed supervised fine-tuning on it using the **ReMUD** dataset, and employed the budget forcing [30] method to enhance its reasoning ability.

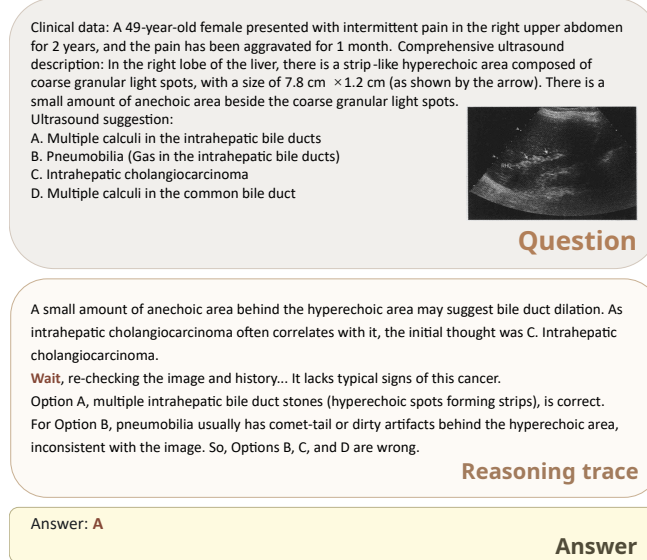


Figure 4: **Budget-forcing with Qwen2.5-VL-Ultra-7B.** Add the marker "Wait, " to the chain of thought to prompt the model to think again. After the model had incorrect thinking, it thought again and corrected the diagnostic result.

215 4.1 Budget forcing

216 Inspired by the innovative approach of s1 [30], we integrate the concept of budget forcing into our
 217 model’s inference process. This strategic decision aims to optimize the model’s performance and
 218 output quality. Budget forcing, in our context, focuses on regulating the length of the chain of
 219 thoughts, which is crucial for maintaining the model’s efficiency and the coherence of its reasoning.

220 More specifically, we target the content enclosed within the `<think></think>` tags. These tags serve
 221 as markers for the model’s internal thought process representation. By controlling the length of this
 222 content, we can ensure that the model’s reasoning neither becomes too concise to be meaningful
 223 nor too verbose to be practical. When the content within these tags is too short, it might indicate
 224 that the model has not fully explored the problem. In such cases, we introduce a "Wait" tag. This
 225 "Wait" acts as a trigger for the model to continue processing, allowing it to generate more in-depth
 226 and comprehensive output. Conversely, if the content is too long, it could lead to inefficiencies and
 227 potential over-thinking. To address this, once the content exceeds the predefined length limit, we
 228 truncate it. After truncation, we add the `</think>` tag as a closure to maintain the integrity of the
 229 tag structure. This ensures that the model’s thought representation remains well-formed and can be
 230 properly processed in subsequent operations. See Fig 4.

231 4.2 Mixed supervised fine-tuning

232 Different from LLaVA-Ultra [16], a ultrasound multimodal large language model that trained
 233 through ultrasound concept feature alignment, visual enhancement and adaptive sampling for data re-
 234 dundancy, we adopt DeepSeek-r1’s distillation model training approach. It makes the training more
 235 efficiency and resource-saving. **ReMUD** encompasses both reasoning and non-reasoning data, with
 236 reasoning data having `<think></think>` tags to represent the thinking trace. Similar to DeepSeek-
 237 R1’s distillation model and s1K, we fine-tune the model via supervised fine-tuning to enhance its
 238 performance, enabling non-reasoning dialogue capabilities and ultrasound disease diagnostic infer-
 239 ence. Additionally, the trained loss function acts on the `<think></think>` tag to optimize the model’s
 240 internal reasoning process for better handling of ultrasound-related diagnosis.

241 4.3 Evaluation metrics

242 In order to make the test results of the model more stable, and in order to fully utilize the chain of
243 thought to explore more possibilities, we chose the pass@1 [8] metric to evaluate our model:

$$\text{pass@1} = \frac{1}{k} \sum_{i=1}^k p_i,$$

244 where k denotes the total number of responses generated for each question, p_i stands for the cor-
245 rectness of the i -th response, typically taking a binary value (1 for correct and 0 for incorrect). By
246 summing these p_i values over the range from $i = 1$ to k and then dividing by k , we obtain the
247 pass@1 value, which represents the average proportion of correct responses among the generated
248 set. Our test dataset is in the form of multi-choice question, so it is easily to check the answer.

249 5 Experiments

250 5.1 Setup

251 We conduct supervised finetuning on the Qwen2.5-VL-7B-Instruct model leveraging ReMUD, and
252 utilize LLaMA-Factory to train our model, Qwen2.5-VL-Ultra-7B. The learning rate is configured
253 as $3e-5$, and we adopt cosine learning rate decay. For model training, the AdamW optimizer is
254 employed. The entire finetuning procedure consists of 3 epochs. All experiments were executed on
255 a single NVIDIA A800 (80G) GPU.

256 5.2 Test data

257 Due to the lack of publicly available ultrasound benchmarks, we have created four different types
258 of test datasets for ReMUD to evaluate the ultrasound-specific capabilities of MLLM models. The
259 descriptions of the specific evaluation datasets are as follows:

- 260 • **USQA-Knowledge.** USQA-Knowledge is automatically generated from ultrasound books,
261 guidelines, and papers. It only contains text information. Specifically, ultrasound knowl-
262 edge is sent to GPT-4o, which is required to generate text-based ultrasound multiple-choice
263 questions based on the provided content, and the answers should be found within the given
264 content. A total of 386 questions were generated, and after manual screening, 353 questions
265 were obtained.
- 266 • **UVQA-Diagnosis.** UVQA-Diagnosis is compiled based on the publicly available dataset,
267 Chinese Ultrasound Report Dataset [24], which contains $7k+$ ultrasound reports with $30k+$
268 images. By utilizing the information provided in the ultrasound reports, we have trans-
269 formed the ultrasound diagnoses and the observed contents into the form of text-and-image
270 single-choice questions, which are used to evaluate the model’s capabilities in diagnosing
271 ultrasound reports. UVQA-Diagnosis contains a total of 386 single-choice text-and-image
272 questions.

273 5.3 Evaluation

274 We evaluate Qwen2.5-VL-Ultra-7B’s ultrasound capability on USQA-knowledge and UVQA-
275 Diagnosis. In order to ensure the randomness of the thinking trace, we choose the pass@1 metric to
276 evaluate the model by setting temperature value of 0.6 and a top- p value of 0.7 to generate $k = 4$
277 responses for each question. In contrast, we compare our model with GPT-4o, Gemini-2.0-Flash-
278 Thinking-Exp, Gemini-2.5-Pro-Exp, Claude-3.7-sonnet-thinking, Qwen2.5-VL-7B-Instruct. See Ta-
279 ble 1, Qwen2.5-VL-Ultra-7B (ours) achieves better performance than other models.

280 5.4 Test-time scaling

281 Qwen2.5-VL-Ultra-7B is a reasoning model. Inspired by s1, we make a experiment by using budget
282 forcing to validate test-time scaling. We forced the addition of the "wait" tag to the parts where
283 the length of the chain of thought was insufficient, enabling the model to think again. It was found

Table 1: Comparison between different MLLMs. Qwen2.5-VL-Ultra-7B achieves better results than other models. DeepSeek-R1 is a single-modal large language model, so it cannot be tested on UVQA-Diagnosis. UVQA-Diagnosis test dataset has three ultrasonic parts: breast, liver and thyroid. "w/o BF" means without using budget forcing.

Model	USQA-Knowledge	UVQA-Diagnosis			
		breast	liver	thyroid	all
Qwen2.5-VL-7B-Instruct	63.1	53.8	52.3	50.9	52.6
GPT-4o	78.3	62.9	62.7	62.8	62.7
Gemini-2.0-Flash-Thinking-Exp	70.8	68.3	66.2	54.9	62.4
Gemini-2.5-Pro-Exp	78.1	70.9	70.7	58.1	66.6
DeepSeek-R1	79.2	-	-	-	Nan
Claude-3.7-sonnet-thinking	66.8	44.1	38.2	37.8	40.2
Qwen2.5-VL-Ultra-7B w/o BF	78.5	90.2	91.5	86.0	88.1
Qwen2.5-VL-Ultra-7B (ours)	80.1	91.8	93.3	88.4	90.1

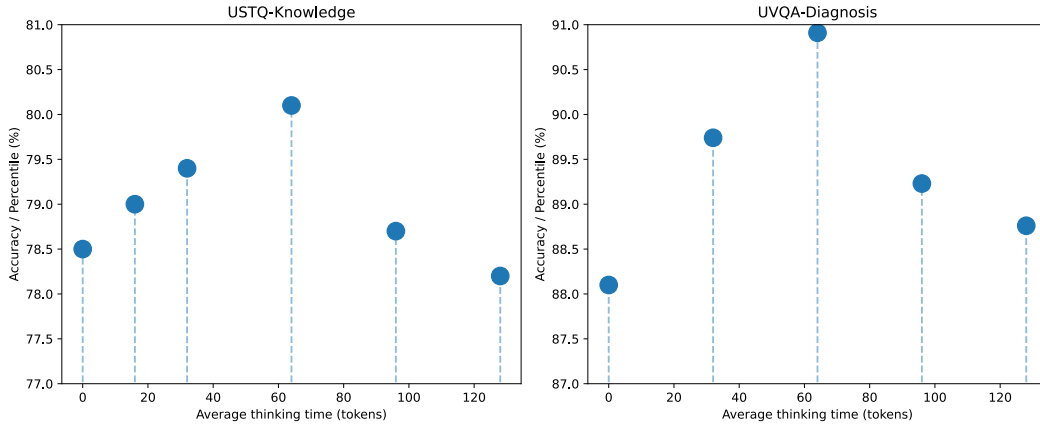


Figure 5: Test-time scaling with Qwen2.5-VL-Ultra-7B. We evaluate Qwen2.5-VL-Ultra-7B on UVQA-Diagnosis using pass@1 and vary test-time compute.

that compared to cases without a chain of thought, the model’s accuracy could be improved when a chain of thought was added. As the chain of thought lengthened, the accuracy first increased and then decreased, see Fig 5.

6 Conclusion

In this study, we successfully developed a pipeline for generating specific domain image-text data, which is crucial for enhancing the performance of MLLMs in specialized fields. By leveraging professional books and papers in the medical ultrasound domain, we created the ReMUD dataset, which significantly improves the performance of the fine-tuned Qwen2.5-VL-Ultra-7B model in medical ultrasound tasks compared to general-domain MLLMs. The open-sourcing of our dataset, codebase, and model parameters will promote further research in specific domain and medical ultrasound fields. Future work could explore applying this pipeline to more specific domains, improving the quality of generated data, and enhancing the reasoning capabilities of models. Additionally, more efficient test-time scaling methods can be investigated to optimize the performance of models during inference.

References

- [1] Google. gemini 2.0 flash thinking mode (gemini-2.0-flash-thinking-exp-1219), december 2024. URL <https://cloud.google.com/vertex-ai/generative-ai/docs/thinking-mode>. 3, 5

- [2] Openai. learning to reason with llms, september 2024. URL <https://openai.com/index/learning-to-reason-with-llms/>. 3
- [3] R. AlSaad, A. Abd-Alrazaq, S. Boughorbel, A. Ahmed, M.-A. Renault, R. Damseh, and J. Sheikh. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of medical Internet research*, 26:e59505, 2024. 1
- [4] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-vl technical report, 2025. URL <https://arxiv.org/abs/2502.13923>. 2, 6
- [5] Z. Bai, P. Wang, T. Xiao, T. He, Z. Han, Z. Zhang, and M. Z. Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024. 1
- [6] B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré, and A. Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. 3
- [7] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, L. Baraldi, M. Cornia, and R. Cucchiara. The revolution of multimodal large language models: A survey, 2024. URL <https://arxiv.org/abs/2402.12451>. 1
- [8] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. de Oliveira Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, A. Ray, R. Puri, G. Krueger, M. Petrov, H. Khlaaf, G. Sastry, P. Mishkin, B. Chan, S. Gray, N. Ryder, M. Pavlov, A. Power, L. Kaiser, M. Bavarian, C. Winter, P. Tillet, F. P. Such, D. Cummings, M. Plappert, F. Chantzis, E. Barnes, A. Herbert-Voss, W. H. Guss, A. Nichol, A. Paino, N. Tezak, J. Tang, I. Babuschkin, S. Balaji, S. Jain, W. Saunders, C. Hesse, A. N. Carr, J. Leike, J. Achiam, V. Misra, E. Morikawa, A. Radford, M. Knight, M. Brundage, M. Murati, K. Mayer, P. Welinder, B. McGrew, D. Amodei, S. McCandlish, I. Sutskever, and W. Zaremba. Evaluating large language models trained on code, 2021. URL <https://arxiv.org/abs/2107.03374>. 8
- [9] Y. Chen, C. Zhang, L. Liu, C. Feng, C. Dong, Y. Luo, and X. Wan. Uscl: Pretraining deep ultrasound image diagnosis model through video contrastive representation learning, 2021. URL <https://arxiv.org/abs/2011.13066>. 5
- [10] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL <https://lmsys.org/blog/2023-03-30-vicuna/>. 2
- [11] DeepSeek-AI, D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, X. Zhang, X. Yu, Y. Wu, Z. F. Wu, Z. Gou, Z. Shao, Z. Li, Z. Gao, A. Liu, B. Xue, B. Wang, B. Wu, B. Feng, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Ding, H. Xin, H. Gao, H. Qu, H. Li, J. Guo, J. Li, J. Wang, J. Chen, J. Yuan, J. Qiu, J. Li, J. L. Cai, J. Ni, J. Liang, J. Chen, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Zhao, L. Wang, L. Zhang, L. Xu, L. Xia, M. Zhang, M. Zhang, M. Tang, M. Li, M. Wang, M. Li, N. Tian, P. Huang, P. Zhang, Q. Wang, Q. Chen, Q. Du, R. Ge, R. Zhang, R. Pan, R. Wang, R. J. Chen, R. L. Jin, R. Chen, S. Lu, S. Zhou, S. Chen, S. Ye, S. Wang, S. Yu, S. Zhou, S. Pan, S. S. Li, S. Zhou, S. Wu, S. Ye, T. Yun, T. Pei, T. Sun, T. Wang, W. Zeng, W. Zhao, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, W. L. Xiao, W. An, X. Liu, X. Wang, X. Chen, X. Nie, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yang, X. Li, X. Su, X. Lin, X. Q. Li, X. Jin, X. Shen, X. Chen, X. Sun, X. Wang, X. Song, X. Zhou, X. Wang, X. Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. Zhang, Y. Xu, Y. Li, Y. Zhao, Y. Sun, Y. Wang, Y. Yu, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Ou, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Xiong, Y. Luo, Y. You, Y. Liu, Y. Zhou, Y. X. Zhu, Y. Xu, Y. Huang, Y. Li, Y. Zheng, Y. Zhu, Y. Ma, Y. Tang, Y. Zha, Y. Yan, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Xie, Z. Zhang, Z. Hao, Z. Ma, Z. Yan, Z. Wu, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Pan, Z. Huang, Z. Xu, Z. Zhang, and Z. Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL <https://arxiv.org/abs/2501.12948>. 3, 5

- [12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2
- [13] S. Doveh, S. Perek, M. J. Mirza, W. Lin, A. Alfassy, A. Arbelle, S. Ullman, and L. Karlin-sky. Towards multimodal in-context learning for vision & language models. *arXiv preprint arXiv:2403.12736*, 2024. 2
- [14] Y. Du, C. Li, R. Guo, C. Cui, W. Liu, J. Zhou, B. Lu, Y. Yang, Q. Liu, X. Hu, D. Yu, and Y. Ma. Pp-ocrv2: Bag of tricks for ultra lightweight ocr system, 2021. URL <https://arxiv.org/abs/2109.03144>. 4
- [15] Z. Gao, B. Niu, X. He, H. Xu, H. Liu, A. Liu, X. Hu, and L. Wen. Interpretable contrastive monte carlo tree search reasoning, 2024. URL <https://arxiv.org/abs/2410.01707>. 3
- [16] X. Guo, W. Chai, S.-Y. Li, and G. Wang. Llava-ultra: Large chinese language and vision assistant for ultrasound, 2024. URL <https://arxiv.org/abs/2410.15074>. 2, 7
- [17] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020. URL <https://arxiv.org/abs/2003.10286>. 2, 6
- [18] X. Hu, L. Gu, Q. An, M. Zhang, L. Liu, K. Kobayashi, T. Harada, R. M. Summers, and Y. Zhu. Expert knowledge-aware image difference graph representation learning for difference-aware medical visual question answering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4156–4165, 2023. 2
- [19] R. Irvine, D. Boubert, V. Raina, A. Liusie, Z. Zhu, V. Mudupalli, A. Korshuk, Z. Liu, F. Cremer, V. Assassi, C.-C. Beauchamp, X. Lu, T. Rialan, and W. Beauchamp. Rewarding chatbots for real-world engagement with millions of users, 2023. URL <https://arxiv.org/abs/2303.06135>. 3
- [20] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. URL <https://arxiv.org/abs/1612.06890>. 2
- [21] N. Levi. A simple model of inference scaling laws, 2024. URL <https://arxiv.org/abs/2410.16377>. 3
- [22] C. Li, W. Liu, R. Guo, X. Yin, K. Jiang, Y. Du, Y. Du, L. Zhu, B. Lai, X. Hu, D. Yu, and Y. Ma. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system, 2022. URL <https://arxiv.org/abs/2206.03001>. 4
- [23] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL <https://arxiv.org/abs/2301.12597>. 2
- [24] J. Li, T. Su, B. Zhao, F. Lv, Q. Wang, N. Navab, Y. Hu, and Z. Jiang. Ultrasound report generation with cross-modality feature alignment via unsupervised guidance, 2024. URL <https://arxiv.org/abs/2406.00644>. 5, 6, 8
- [25] Z. Liang, Y. Xu, Y. Hong, P. Shang, Q. Wang, Q. Fu, and K. Liu. A survey of multimodal large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409, 2024. 1
- [26] W. Lin, Z. Zhao, X. Zhang, C. Wu, Y. Zhang, Y. Wang, and W. Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *arXiv preprint arXiv:2303.07240*, 2023. 6
- [27] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2
- [28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

- [29] H. Luo, Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, Q. Lin, S. Chen, Y. Tang, and D. Zhang. Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct, 2025. URL <https://arxiv.org/abs/2308.09583>. 3
- [30] N. Muennighoff, Z. Yang, W. Shi, X. L. Li, L. Fei-Fei, H. Hajishirzi, L. Zettlemoyer, P. Liang, E. Candès, and T. Hashimoto. s1: Simple test-time scaling, 2025. URL <https://arxiv.org/abs/2501.19393>. 6, 7
- [31] OpenAI, :, A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Mdry, A. Baker-Whitcomb, A. Beutel, A. Borzunov, A. Carney, A. Chow, A. Kirillov, A. Nichol, A. Paino, A. Renzin, A. T. Passos, A. Kirillov, A. Christakis, A. Conneau, A. Kamali, A. Jabri, A. Moyer, A. Tam, A. Crookes, A. Tootoochian, A. Tootoonchian, A. Kumar, A. Vallone, A. Karpathy, A. Braunstein, A. Cann, A. Codisoti, A. Galu, A. Kondrich, A. Tulloch, A. Mishchenko, A. Baek, A. Jiang, A. Pelisse, A. Woodford, A. Gosalia, A. Dhar, A. Pantuliano, A. Nayak, A. Oliver, B. Zoph, B. Ghorbani, B. Leimberger, B. Rossen, B. Sokolowsky, B. Wang, B. Zweig, B. Hoover, B. Samic, B. McGrew, B. Spero, B. Giertler, B. Cheng, B. Lightcap, B. Walkin, B. Quinn, B. Guaraci, B. Hsu, B. Kellogg, B. Eastman, C. Lugaresi, C. Wainwright, C. Bassin, C. Hudson, C. Chu, C. Nelson, C. Li, C. J. Shern, C. Conger, C. Barette, C. Voss, C. Ding, C. Lu, C. Zhang, C. Beaumont, C. Hallacy, C. Koch, C. Gibson, C. Kim, C. Choi, C. McLeavey, C. Hesse, C. Fischer, C. Winter, C. Czarnecki, C. Jarvis, C. Wei, C. Koumouzelis, D. Sherburn, D. Kappler, D. Levin, D. Levy, D. Carr, D. Farhi, D. Mely, D. Robinson, D. Sasaki, D. Jin, D. Valladares, D. Tsipras, D. Li, D. P. Nguyen, D. Findlay, E. Oiwoh, E. Wong, E. Asdar, E. Proehl, E. Yang, E. Antonow, E. Kramer, E. Peterson, E. Sigler, E. Wallace, E. Brevdo, E. Mays, F. Khorasani, F. P. Such, F. Raso, F. Zhang, F. von Lohmann, F. Sulit, G. Goh, G. Oden, G. Salmon, G. Starace, G. Brockman, H. Salman, H. Bao, H. Hu, H. Wong, H. Wang, H. Schmidt, H. Whitney, H. Jun, H. Kirchner, H. P. de Oliveira Pinto, H. Ren, H. Chang, H. W. Chung, I. Kivlichan, I. O’Connell, I. O’Connell, I. Osband, I. Silber, I. Sohl, I. Okuyucu, I. Lan, I. Kostrikov, I. Sutskever, I. Kanitscheider, I. Gulrajani, J. Coxon, J. Menick, J. Pachocki, J. Aung, J. Betker, J. Crooks, J. Lennon, J. Kiros, J. Leike, J. Park, J. Kwon, J. Phang, J. Teplitz, J. Wei, J. Wolfe, J. Chen, J. Harris, J. Varavva, J. G. Lee, J. Shieh, J. Lin, J. Yu, J. Weng, J. Tang, J. Yu, J. Jang, J. Q. Candela, J. Beutler, J. Landers, J. Parish, J. Heidecke, J. Schulman, J. Lachman, J. McKay, J. Uesato, J. Ward, J. W. Kim, J. Huizinga, J. Sitkin, J. Kraaijeveld, J. Gross, J. Kaplan, J. Snyder, J. Achiam, J. Jiao, J. Lee, J. Zhuang, J. Harri-man, K. Fricke, K. Hayashi, K. Singhal, K. Shi, K. Karthik, K. Wood, K. Rimbach, K. Hsu, K. Nguyen, K. Gu-Lemberg, K. Button, K. Liu, K. Howe, K. Muthukumar, K. Luther, L. Ahmad, L. Kai, L. Itow, L. Workman, L. Pathak, L. Chen, L. Jing, L. Guy, L. Fedus, L. Zhou, L. Mamitsuka, L. Weng, L. McCallum, L. Held, L. Ouyang, L. Feuvrier, L. Zhang, L. Kondraciuk, L. Kaiser, L. Hewitt, L. Metz, L. Doshi, M. Aflak, M. Simens, M. Boyd, M. Thompson, M. Dukhan, M. Chen, M. Gray, M. Hudnall, M. Zhang, M. Aljube, M. Litwin, M. Zeng, M. Johnson, M. Shetty, M. Gupta, M. Shah, M. Yatbaz, M. J. Yang, M. Zhong, M. Glaese, M. Chen, M. Janner, M. Lampe, M. Petrov, M. Wu, M. Wang, M. Fradin, M. Pokrass, M. Castro, M. O. T. de Castro, M. Pavlov, M. Brundage, M. Wang, M. Khan, M. Murati, M. Bavarian, M. Lin, M. Yesildal, N. Soto, N. Gimelshein, N. Cone, N. Staudacher, N. Summers, N. La-Fontaine, N. Chowdhury, N. Ryder, N. Stathas, N. Turley, N. Tezak, N. Felix, N. Kudige, N. Keskar, N. Deutsch, N. Bundick, N. Puckett, O. Nachum, O. Okelola, O. Boiko, O. Murk, O. Jaffe, O. Watkins, O. Godement, O. Campbell-Moore, P. Chao, P. McMillan, P. Belov, P. Su, P. Bak, P. Bakkum, P. Deng, P. Dolan, P. Hoeschele, P. Welinder, P. Tillet, P. Pronin, P. Tillet, P. Dhariwal, Q. Yuan, R. Dias, R. Lim, R. Arora, R. Troll, R. Lin, R. G. Lopes, R. Puri, R. Miyara, R. Leike, R. Gaubert, R. Zamani, R. Wang, R. Donnelly, R. Honsby, R. Smith, R. Sahai, R. Ramchandani, R. Huet, R. Carmichael, R. Zellers, R. Chen, R. Chen, R. Nig-matullin, R. Cheu, S. Jain, S. Altman, S. Schoenholz, S. Toizer, S. Miserendino, S. Agarwal, S. Culver, S. Ethersmith, S. Gray, S. Grove, S. Metzger, S. Hermani, S. Jain, S. Zhao, S. Wu, S. Jomoto, S. Wu, Shuaiqi, Xia, S. Phene, S. Papay, S. Narayanan, S. Coffey, S. Lee, S. Hall, S. Balaji, T. Broda, T. Stramer, T. Xu, T. Gogineni, T. Christianson, T. Sanders, T. Patwardhan, T. Cunninghamman, T. Degry, T. Dimson, T. Raoux, T. Shadwell, T. Zheng, T. Underwood, T. Markov, T. Sherbakov, T. Rubin, T. Stasi, T. Kaftan, T. Heywood, T. Peterson, T. Walters, T. Eloundou, V. Qi, V. Moeller, V. Monaco, V. Kuo, V. Fomenko, W. Chang, W. Zheng, W. Zhou, W. Manassra, W. Sheu, W. Zaremba, Y. Patil, Y. Qian, Y. Kim, Y. Cheng,

461 Y. Zhang, Y. He, Y. Zhang, Y. Jin, Y. Dai, and Y. Malkov. Gpt-4o system card, 2024. URL
462 <https://arxiv.org/abs/2410.21276>. 1, 3

463 [32] Y. Qian, H. Ye, J.-P. Fauconnier, P. Gräsch, Y. Yang, and Z. Gan. Mia-bench: Towards better
464 instruction following evaluation of multimodal llms. *arXiv preprint arXiv:2407.01509*, 2024.
465 2

466 [33] Qwen, :, A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang,
467 H. Wei, H. Lin, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Lin, K. Dang, K. Lu,
468 K. Bao, K. Yang, L. Yu, M. Li, M. Xue, P. Zhang, Q. Zhu, R. Men, R. Lin, T. Li, T. Tang,
469 T. Xia, X. Ren, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Wan, Y. Liu, Z. Cui, Z. Zhang, and Z. Qiu.
470 Qwen2.5 technical report, 2025. URL <https://arxiv.org/abs/2412.15115>. 2

471 [34] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell,
472 P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from
473 natural language supervision, 2021. URL <https://arxiv.org/abs/2103.00020>. 2

474 [35] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R.
475 Bowman. Gpqa: A graduate-level google-proof qa benchmark, 2023. URL <https://arxiv.org/abs/2311.12022>. 2

476

477 [36] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal. Ai models
478 collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024. 2

479 [37] C. Snell, J. Lee, K. Xu, and A. Kumar. Scaling llm test-time compute optimally can be more ef-
480 fective than scaling model parameters, 2024. URL <https://arxiv.org/abs/2408.03314>.
481 3

482 [38] S. Song, X. Li, S. Li, S. Zhao, J. Yu, J. Ma, X. Mao, and W. Zhang. How to bridge the
483 gap between modalities: Survey on multimodal large language model, 2025. URL <https://arxiv.org/abs/2311.07594>. 1
484

485 [39] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovi-
486 cova, A. Ramé, M. Rivière, L. Rouillard, T. Mesnard, G. Cideron, J. bastien Grill, S. Ramos,
487 E. Yvinec, M. Casbon, E. Pot, I. Penchev, G. Liu, F. Visin, K. Kenealy, L. Beyer, X. Zhai,
488 A. Tsitsulin, R. Busa-Fekete, A. Feng, N. Sachdeva, B. Coleman, Y. Gao, B. Mustafa, I. Barr,
489 E. Parisotto, D. Tian, M. Eyal, C. Cherry, J.-T. Peter, D. Sinopalnikov, S. Bhupatiraju, R. Agar-
490 wal, M. Kazemi, D. Malkin, R. Kumar, D. Vilar, I. Brusilovsky, J. Luo, A. Steiner, A. Friesen,
491 A. Sharma, A. Sharma, A. M. Gilady, A. Goedeckemeyer, A. Saade, A. Feng, A. Kolesnikov,
492 A. Bendebury, A. Abdagic, A. Vadi, A. György, A. S. Pinto, A. Das, A. Bapna, A. Miech,
493 A. Yang, A. Paterson, A. Shenoy, A. Chakrabarti, B. Piot, B. Wu, B. Shahriari, B. Petrini,
494 C. Chen, C. L. Lan, C. A. Choquette-Choo, C. Carey, C. Brick, D. Deutsch, D. Eisen-
495 bud, D. Cattle, D. Cheng, D. Paparas, D. S. Sreepathihalli, D. Reid, D. Tran, D. Zelle,
496 E. Noland, E. Huizenga, E. Kharitonov, F. Liu, G. Amirkhanyan, G. Cameron, H. Hashemi,
497 H. Klimczak-Pluciska, H. Singh, H. Mehta, H. T. Lehri, H. Hazimeh, I. Ballantyne, I. Szpek-
498 tor, I. Nardini, J. Pouget-Abadie, J. Chan, J. Stanton, J. Wieting, J. Lai, J. Orbay, J. Fernandez,
499 J. Newlan, J. yeong Ji, J. Singh, K. Black, K. Yu, K. Hui, K. Vodrahalli, K. Greff, L. Qiu,
500 M. Valentine, M. Coelho, M. Ritter, M. Hoffman, M. Watson, M. Chaturvedi, M. Moyni-
501 han, M. Ma, N. Babar, N. Noy, N. Byrd, N. Roy, N. Momchev, N. Chauhan, N. Sachdeva,
502 O. Bunyan, P. Botarda, P. Caron, P. K. Rubenstein, P. Culliton, P. Schmid, P. G. Sessa,
503 P. Xu, P. Stanczyk, P. Tafti, R. Shivanna, R. Wu, R. Pan, R. Rokni, R. Willoughby, R. Vallu,
504 R. Mullins, S. Jerome, S. Smoot, S. Girgin, S. Iqbal, S. Reddy, S. Sheth, S. Pöder, S. Bhat-
505 nagar, S. R. Panyam, S. Eiger, S. Zhang, T. Liu, T. Yacovone, T. Liechty, U. Kalra, U. Evci,
506 V. Misra, V. Roseberry, V. Feinberg, V. Kolesnikov, W. Han, W. Kwon, X. Chen, Y. Chow,
507 Y. Zhu, Z. Wei, Z. Egyed, V. Cotruta, M. Giang, P. Kirk, A. Rao, K. Black, N. Babar, J. Lo,
508 E. Moreira, L. G. Martins, O. Sanseviero, L. Gonzalez, Z. Gleicher, T. Warkentin, V. Mir-
509 rokni, E. Senter, E. Collins, J. Barral, Z. Ghahramani, R. Hadsell, Y. Matias, D. Sculley,
510 S. Petrov, N. Fiedel, N. Shazeer, O. Vinyals, J. Dean, D. Hassabis, K. Kavukcuoglu, C. Farabet,
511 E. Buchatskaya, J.-B. Alayrac, R. Anil, Dmitry, Lepikhin, S. Borgeaud, O. Bachem, A. Joulin,
512 A. Andreev, C. Hardin, R. Dadashi, and L. Hussenot. Gemma 3 technical report, 2025. URL
513 <https://arxiv.org/abs/2503.19786>. 2

- [40] K. Team, A. Du, B. Yin, B. Xing, B. Qu, B. Wang, C. Chen, C. Zhang, C. Du, C. Wei, C. Wang, D. Zhang, D. Du, D. Wang, E. Yuan, E. Lu, F. Li, F. Sung, G. Wei, G. Lai, H. Zhu, H. Ding, H. Hu, H. Yang, H. Zhang, H. Wu, H. Yao, H. Lu, H. Wang, H. Gao, H. Zheng, J. Li, J. Su, J. Wang, J. Deng, J. Qiu, J. Xie, J. Wang, J. Liu, J. Yan, K. Ouyang, L. Chen, L. Sui, L. Yu, M. Dong, M. Dong, N. Xu, P. Cheng, Q. Gu, R. Zhou, S. Liu, S. Cao, T. Yu, T. Song, T. Bai, W. Song, W. He, W. Huang, W. Xu, X. Yuan, X. Yao, X. Wu, X. Zu, X. Zhou, X. Wang, Y. Charles, Y. Zhong, Y. Li, Y. Hu, Y. Chen, Y. Wang, Y. Liu, Y. Miao, Y. Qin, Y. Chen, Y. Bao, Y. Wang, Y. Kang, Y. Liu, Y. Du, Y. Wu, Y. Wang, Y. Yan, Z. Zhou, Z. Li, Z. Jiang, Z. Zhang, Z. Yang, Z. Huang, Z. Huang, Z. Zhao, Z. Chen, and Z. Lin. Kimi-vl technical report, 2025. URL <https://arxiv.org/abs/2504.07491>. 2
- [41] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. Llama: Open and efficient foundation language models, 2023. URL <https://arxiv.org/abs/2302.13971>. 2
- [42] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang, Z. Xie, Y. Wu, K. Hu, J. Wang, Y. Sun, Y. Li, Y. Piao, K. Guan, A. Liu, X. Xie, Y. You, K. Dong, X. Yu, H. Zhang, L. Zhao, Y. Wang, and C. Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL <https://arxiv.org/abs/2412.10302>. 2
- [43] Y. Xie, K. Kawaguchi, Y. Zhao, X. Zhao, M.-Y. Kan, J. He, and Q. Xie. Self-evaluation guided beam search for reasoning, 2023. URL <https://arxiv.org/abs/2305.00633>. 3
- [44] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan. Qwen2 technical report, 2024. URL <https://arxiv.org/abs/2407.10671>. 2
- [45] Z. Yang, N. Band, S. Li, E. Candès, and T. Hashimoto. Synthetic continued pretraining, 2024. URL <https://arxiv.org/abs/2409.07431>. 3
- [46] L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller, and W. Liu. Metamath: Bootstrap your own mathematical questions for large language models, 2024. URL <https://arxiv.org/abs/2309.12284>. 3
- [47] C. Zhang, K. Lin, Z. Yang, J. Wang, L. Li, C.-C. Lin, Z. Liu, and L. Wang. Mm-narrator: Narrating long-form videos with multimodal in-context learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13647–13657, 2024. 2
- [48] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola. Multimodal chain-of-thought reasoning in language models, 2024. URL <https://arxiv.org/abs/2302.00923>. 2
- [49] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, Z. Gao, E. Cui, X. Wang, Y. Cao, Y. Liu, X. Wei, H. Zhang, H. Wang, W. Xu, H. Li, J. Wang, N. Deng, S. Li, Y. He, T. Jiang, J. Luo, Y. Wang, C. He, B. Shi, X. Zhang, W. Shao, J. He, Y. Xiong, W. Qu, P. Sun, P. Jiao, H. Lv, L. Wu, K. Zhang, H. Deng, J. Ge, K. Chen, L. Wang, M. Dou, L. Lu, X. Zhu, T. Lu, D. Lin, Y. Qiao, J. Dai, and W. Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL <https://arxiv.org/abs/2504.10479>. 2

559 A Prompts

Prompting GPT-4 to generate VQA instruction-following data

```
messages = [{"role": "system", "content": """"You're an ultrasound medicine AI assistant.  
Given a page image from an ultrasound medicine book and a corresponding JSON list:
```

```
- Generate multi-round Q&A dialogues for each ultrasound image in the JSON based on "caption" and relevant  
text, in ShareGPT format, without using image numbers.  
- Standardize the output into JSON format, specifically the ShareGPT format, for SFT fine-tuning, that  
is, '''json\n[{"conversations": [{"from": "human", "value": "Question"}, {"from": "gpt", "value": "Answer"}]}] ,  
{"conversations": [{"from": "human", "value": "Question"}, {"from": "gpt", "value": "Answer"}]}]\n"  
- The ShareGPT format of the output should correspond to the JSON list, that is, the first element of the output  
list corresponds to the first element of the JSON list you received. Then, it can include multiple rounds of  
dialogues, and it should meet the ShareGPT format.  
- The elements of the JSON list required to be output have the same number as the elements of the JSON list  
given to you, corresponding one by one, that is, each ultrasonic image corresponds to a dialogue in the multi-  
round ShareGPT format.  
""",  
}]
```

```
for sample in fewshot_samples:
```

```
    messages.append({"role": "user", "content": sample['context']})  
    messages.append({"role": "assistant", "content": sample['response']})
```

```
messages.append({"role": "user", "content": query})
```

Figure 6: Messages we use to prompt GPT-4o to generate VQA instruction-following data. Different adjustments to the prompts may be made when handling different data. Please see Fig 2 for one of the few-shot examples.