

ICT-QA: Question Answering over Multi-modal Contexts including Image, Chart, and Text Modalities

Youngrok Jang* Hyesoo Kong* Gyeonghun Kim Yejin Lee Jungkyu Choi Kyunghoon Bae
LG AI Research
Seoul, South Korea

{jyrok3357, hyesoo.kong, ghkayne.kim, yejin.lee, stanleyjk.choi, k.bae}@lgresearch.ai

Abstract

For question answering in multi-modal contexts that include image, chart, and text modalities, a model must be proficient in understanding each individual modality. Furthermore, the model must be able to find the necessary evidence from multiple modalities and generate answers through cross-modal reasoning for some questions. In this paper, we propose the Image and Chart Instruction Tuning (IC-tuning) method to enhance the model’s comprehension of each modality. Specifically, we introduce visual-aware chart instruction-following data that describe both precise numerical values and visual information on the charts. We then train a Large Language Model (LLM) with a model architecture that utilizes an image-specific encoder and a chart-specific encoder. Our experiments demonstrate that this method achieves state-of-the-art performance in Chart Summarization and Open-ended Chart question answering (OpenCQA) tasks while having minimal impact on image and language benchmark performance. Although the IC-tuned model shows great comprehension performance for each modality, it still struggles with question answering tasks in multi-modal contexts because it is only trained on data for understanding each individual modality. To address this, we introduce the Question Answering over Image, Chart, and Text (ICT-QA) dataset, designed specifically for question answering in multi-modal contexts. After further training the IC-tuned LLM with the ICT-QA dataset, our evaluations demonstrate that ICT-QA significantly improves the quality of answers for both single-modal questions, where only one modality needs to be referenced from multiple modalities, and cross-modal questions, which require reasoning across multiple modalities.

1. Introduction

Information is conveyed not only through text but also through various other modalities. Many documents incorporate images to provide visual evidence and charts to effectively represent data trends. For example, Pew Research Center¹ publishes articles on public opinion polling, demographic research and other data-driven social science research, incorporating text as well as images and charts. Addressing question answering in these multi-modal contexts presents two significant challenges: (1) The model must be proficient in understanding image, chart, and text modalities. (2) The model must be able to find the necessary evidence from multiple modalities and generate answers through cross-modal reasoning for some questions.

For the challenge (1), there is a variety of research on Multi-modal Large Language Model (MLLM), which can understand image and text modalities [3, 5, 11, 31, 36, 37, 64, 67]. For example, LLaVA [36, 37] connects the image encoder, CLIP [52], to the Large Language Model (LLM) through visual instruction tuning. In the case of chart modality, similar to LLaVA, there are studies that connect CLIP by treating charts as images, or use other chart-specific encoders to connect to LLM [15, 45, 47]. These studies propose chart instruction-following data for various tasks generated by LLMs such as GPT [50].

However, we identify two opportunities to improve the model’s comprehension of all three modalities. First, all the chart instruction-following data proposed in these studies does not consider the visual information on the charts. Their approach uses a table converted from the chart as input to the LLM, resulting in data that includes only numerical information and lacks the visual information on the charts. Second, these studies do not focus on improving the performance across all three modalities. For example, post-tuning LLaVA with chart instruction-following data, as in ChartLlama [15], improves chart comprehension but significantly decreases understanding of image and text modalities due

*Equal Contributions

¹<https://www.pewresearch.org/>

to catastrophic forgetting, as observed in Section 4.2.

Therefore, we propose the *Image and Chart instruction tuning (IC-tuning)* method to enhance the model’s comprehension of each modality as shown in Figure 1. First, we propose generating visual-aware chart instruction-following data by using GPT-4V [50] with both tables and chart images. Note that our proposed visual-aware chart instruction-following data considers both the exact numerical values in the tables and the visual information on the chart images, such as the text around the chart (*e.g.*, title), legend (*e.g.*, x and y axes), and color, as shown in Figure 2. Second, we train a model with a mixture of multi-modal instruction following data that includes existing data and our proposed data, encompassing image, chart, and text modalities. For this, we adopt a model architecture that utilizes CLIP and Unichart [45] encoders separately for each image and chart modality.

Finally, we demonstrate that our IC-tuned LLM is competitive on traditional chart benchmarks and achieves the highest performance on G-Eval [40] in Chart Summarization and Open-ended Chart Question Answering (OpenCQA) [20] tasks. Furthermore, through ablation studies, we show that our model architecture, which employs two separate encoders, is highly effective in improving chart benchmark performance while minimizing the impact on vision and language benchmark performance.

Although our IC-tuned LLM shows excellent performance in each modality, the challenge (2) remains unresolved. As shown in Figure 1, questions in multi-modal contexts include both single-modal questions, requiring finding evidence from a specific modality, and cross-modal questions, requiring reasoning across multiple modalities. As an example of related work, LLaVA-NeXT-Interleave [30] is fine-tuned for multi-image tasks and can handle multi-modal contexts that include both images and charts by treating the chart as another image, similar to a multi-image task. However, it still does not show great performance, as observed in Section 4.3. Furthermore, while there are studies proposing multi-modal question answering data addressing image, table, and text modalities [16, 60], there is no research specifically addressing the image, chart, and text modalities.

Hence, we propose the *Question Answering over Image, Chart, and Text (ICT-QA)* dataset, which includes both single-modal and cross-modal questions in multi-modal contexts. After further training IC-tuned LLM with the ICT-QA dataset and evaluating it with G-Eval, we demonstrate that the model achieves higher-quality answers for multi-modal contexts compared to the only IC-tuned model, LLaVA, and LLaVA-NeXT-Interleave. Additionally, through ablation studies, we show that ICT-QA enhances the capability to find relevant evidence in multi-modal contexts and strengthens cross-modal reasoning.

Our main contributions are: (i) We introduce the IC-tuning method to improve the model’s comprehension of each modality: chart, image, and text. In this method, we propose a visual-aware chart instruction-following dataset and a model architecture that utilizes both an image-specific encoder and a chart-specific encoder. (ii) Our model demonstrates state-of-the-art (SOTA) performance in G-Eval for Chart Summarization and Open-ended Chart Question Answering (OpenCQA) tasks, while having minimal impact on image and language benchmark performance. (iii) Furthermore, we propose the ICT-QA dataset, designed specifically for question answering in multi-modal contexts including image, chart and text modalities. Through experiments, we demonstrate that ICT-QA significantly improves the quality of answers for both single-modal and cross-modal questions within multi-modal contexts.

2. Image and Chart instruction tuning (IC-tuning) method

For question answering in multi-modal contexts that include image, chart, and text modalities, a model must be proficient in understanding each individual modality. Hence, we propose the *Image and Chart Instruction Tuning (IC-tuning)* method to enhance the LLM’s comprehension of each modality, as shown in Figure 1. In the following sections, we first describe the visual-aware chart instruction-following data, and then describe the model architecture and training strategy.

2.1. Visual-aware chart instruction-following data

Chart summarization, open-ended question answering, and reasoning QA are popular tasks for understanding charts [45]. Therefore, we use GPT-4V [50] to generate 57K visual-aware chart instruction-following data for these tasks by utilizing both tables and chart images.² Since GPT-4V can recognize both tables and images, it can generate chart instruction-following data including the exact numerical values in the table and visual information expressed on the chart as shown in Figure 2.

We collect chart and table pairs from the Unichart dataset [45] and ChartQA [44], which consist of various types of charts from diverse sources. Using these pairs, we first generate chart summarizations with GPT-4V. We then generate questions and answers for open-ended QA based on the chart summarizations. As a result, the generated chart summarizations and open-ended questions offer a more holistic comprehension of the chart by incorporating visual information, compared to those generated by using only tables. For reasoning QA, we design it similarly to open-ended QA

²The statistics for generated data can be found in Table 9 of Appendix D, and all the prompts used to generate the data can be found in the Appendix G.

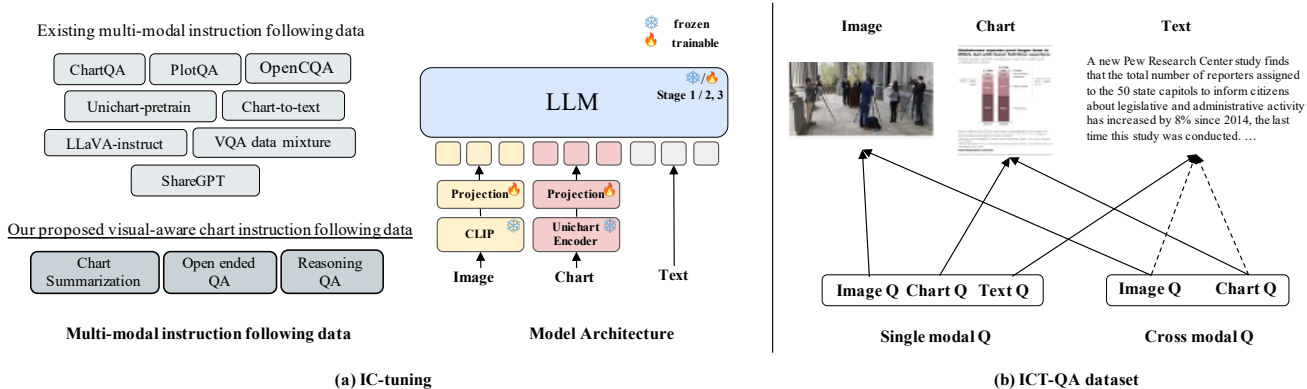


Figure 1. **The overview of IC-tuning and ICT-QA dataset.** In the IC-tuning method shown in Figure (a), we train the LLM using a mixture of multi-modal instruction-following data, incorporating our proposed visual-aware chart instruction-following data. Additionally, we adopt a model architecture that uses CLIP (Radford et al. 2021) and Unichart (Masry et al. 2023) encoders separately for each image and chart modality. Figure (b) represents the ICT-QA dataset, which involves both single-modal and cross-modal questions within multi-modal contexts that incorporate image, chart, and text modalities. Image Q, Chart Q, and Text Q represent questions related to each specific modality, and arrows indicate the modalities that should be referenced to answer the questions. For example, Chart Q, such as “What is this chart trying to express?” may require reasoning across both chart and text modalities. In the illustration, the solid and dashed arrows point to the chart and text modalities, respectively, indicating whether they should be referenced primarily or secondarily.

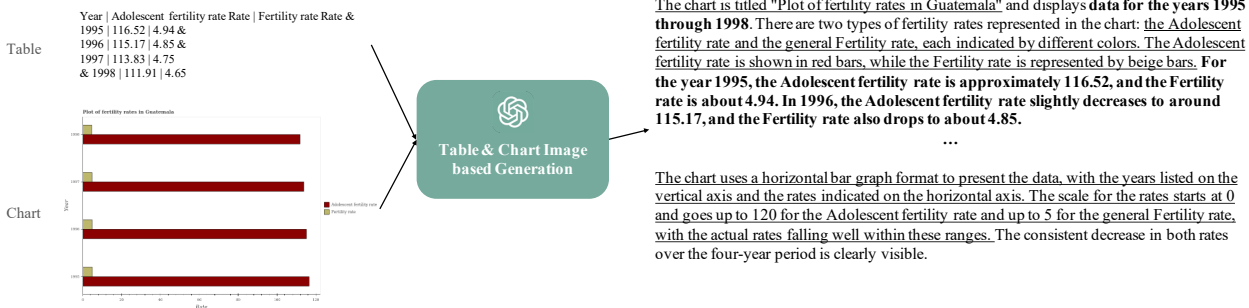


Figure 2. **The example of Chart Summarization data in the visual-aware chart instruction-following data.** Underlined text represents visual descriptions derived from chart image and bold text represents descriptions of numerical values obtained from the table.

but incorporate Chain of Thought (CoT) style answers, akin to Deplot [34].³

2.2. Model Architecture & Training

We adopt a model architecture that uses CLIP [52] and the Unichart encoder [45] separately for each image and chart modality, as shown in Figure 1. The Unichart encoder, a Donut [24] based model fine-tuned on various chart tasks, is better suited for chart encoding compared to CLIP. Additionally, using separate encoders for images and charts ensures that the model is less affected by task interference across different modalities during multi-task tuning on a

³For reasoning QA, we used only the tables and questions from ChartQA without the chart images, as numerical calculation is crucial. We prompted the model to generate CoT-style answers and then filtered out cases where the generated final answers did not match the gold answers from ChartQA.

mixture of multi-modal instruction-following data, as we will discuss in Section 4.2.

IC-tuning consists of a three-stage training process. Note that the details of training data of each stage is described in Appendix D. **Stage 1 is pre-training for image and chart feature alignment.** In this stage, we train only the projection layer of CLIP with image-text pairs and the projection layer of the Unichart encoder with chart-text pairs and chart-table pairs, while keeping the parameters of the LLM frozen. **Stage 2 is fine-tuning for image and chart instruction-following.** In this stage, we train the LLM and all projection layers with a mixture of publicly available existing multi-modal instruction-following data, which include image, chart, and text modalities. Note that the chart instruction-following data used in this stage is generated by LLMs using only tables or by humans. According to the experimental results in Section 4.2, when we proceed to stage

3 after tuning with this data, the model’s chart comprehension performance improves. **Stage 3 is fine-tuning for visual-aware chart instruction-following.** In this stage, we further-train the LLM and all projection layers with our visual-aware chart instruction-following data. To prevent catastrophic forgetting problem, we also sample and use image and text instruction-following data used in stage 2.

3. Question Answering over Image, Chart and Text (ICT-QA) dataset

3.1. Motivation

Through IC-tuning, we enable the model to effectively understand each of the image, chart, and text modalities. However, the model still struggles with QA tasks for multi-modal contexts consisting of image, chart and text modalities, as shown by the examples in Figures 5 and 6 in Appendix B.2. This is due to the following challenges. For single-modal questions, the model must be able to find the evidence for the answer within a multi-modal contexts. For cross-modal questions, the model must be able to find evidence across multiple modalities and generate the answer through cross-modal reasoning.

Hence, we propose *Question Answering over Image, Chart and Text(ICT-QA)* dataset, which includes both single-modal and cross-modal questions for the multi-modal contexts as shown in Figure 1 and Table 1.

3.2. Multi-modal contexts collection and QA generation

We collect 963 articles from the Pew Research Center⁴ for the multi-modal contexts of real-world scenario. These articles are about opinion poll, content analysis and other data-driven social science research and include not only text but also image and chart modalities. We construct a multi-modal contexts utilizing by combining one image, one chart and full text from each crawled article.

In order to generate single-modal and cross-modal questions for multi-modal contexts, we define 9 types of questions (3 types per each modality). Table 1 shows the example of each question type. The target modalities identify where the evidence of the question comes from. *Image + Text* or *Chart + Text* represent cross-modal questions that require reasoning across both image and text, or chart and text, respectively. Note that we do not consider cross-modal reasoning over chart and image, because charts and images are rarely relevant to each other.

We use GPT-4V to generate 7K pairs of predefined types of questions and answers for the collected multi-modal contexts.⁵ For image and chart modality, we generate descrip-

tions of the image and chart, and then generate questions and answers.⁶ Note that when generating questions for image and chart modalities, we also include paragraphs near the image and chart in the prompt to facilitate the generation of cross-modal questions with text.

3.3. Training

Since each training example in the ICT-QA dataset includes both image and chart modalities, we insert special tokens, [Image] or [Chart], before the image and chart modalities to distinguish them. Additionally, we interleave the image or chart at their respective positions within the multi-modal context.

4. Experiment

4.1. Implementation details

For base LLM, similar to LLaVA 1.5 [36], we adopt Vicuna 7B [8], which is Llama 2 [62] fine-tuned with ShareGPT [1]. Refer to Appendix D for training data of each stage and hyper-parameters.

4.2. IC-tuning evaluation

This section describes the evaluation results of IC-tuning method. First, we evaluate chart comprehension performance. Second, we conduct ablation studies to analyze the effectiveness of IC-tuning method.

Traditional chart benchmark evaluation. First, we evaluate our model using traditional chart benchmarks: ChartQA [44], Chart-to-text [21], Chart-to-table [44] and OpenCQA [20]. The metrics we use are relaxed accuracy for ChartQA, BLEU [51] for Chart-to-text, Relative Number Set Similarity (RNSS) [44] and Relative Mapping Similarity (RMS) [34] for Chart-to-table, and BLEU for OpenCQA.

Table 2 shows the evaluation results for traditional chart benchmarks. *Vicuna (IC-tuning)* is our IC-tuned model, which is Vicuna 7B tuned from stage 1 to 3. *Vicuna (IC-tuning)* achieves the highest performance in Chart-to-table and OpenCQA tasks, while demonstrating competitive performance in ChartQA and Chart-to-text tasks.

For ChartQA, ChartAst [47] achieves the highest score of 79.90, surpassing *Vicuna (IC-tuning)*’s score of 74.52. This is likely because ChartAst utilizes a larger 13B language model and generates step-by-step program functions and arguments to calculate answers, similar to the Program-of-Thoughts (PoT) approach [7]. In contrast, *Vicuna (IC-tuning)* is trained on CoT-style Reasoning QA data, relying solely on the language model’s capability rather than program interpreters.

For Chart-to-text, *Vicuna (IC-tuning)* demonstrates competitive performance, though it is not as high as expected.

⁶Appendix H shows the prompts for generation.

⁴<https://www.pewresearch.org/>

⁵Table 10 of Appendix E shows the statistics of generated ICT-QA dataset.

Target modalities	Type	Example
Image	Identification	What is the main object being displayed in the hands of the person in the image?
Image + Text	Contextual relation	How does the image relate to the context provided by the Pew Research Center study on U.S. adult TikTok users?
Image + Text	Purpose	Why is the image included in the document regarding the Pew Research Center study?
Chart	Identification	What percentage of U.S. adult TikTok users reported that they find the videos on their "For You" page "Somewhat" interesting?
Chart	Comparison	How does the percentage of users who find the content "Extremely" interesting compare to those who find it "Not at all" interesting?
Chart + Text	Purpose	Why is the chart included in the document?
Text	Identification	What percentage of U.S. adults on TikTok have never posted a video, and how does this compare to the percentage that has posted videos publicly?
Text	Reasoning	Why might TikTok users who have posted videos be more engaged with the platform compared to those who have not posted?
Text	Summarization	What are the key findings from the Pew Research Center study on how U.S. adults use TikTok?

Table 1. The examples of each question types from ICT-QA

Model	#Params	ChartQA			Chart-to-text		Chart-to-table		OpenCQA
		aug.	human	avg.	Pew	Statista	($RNSS$ RMS_{F1}) ChartQA	($BLEU$) OpenCQA	
Pix2Struct[26]	282M	81.60	30.50	56.00	10.30	38.00	-	-	-
Matcha[35]	282M	90.20	38.20	64.20	-	-	-	-	-
Unichart[45]	201M	88.56	43.92	66.24	12.48	38.21	<u>94.01</u>	91.10	14.88
ChartInstruct[46]	7B	87.76	45.52	66.64	<u>13.83</u>	43.53	-	-	14.78
ChartLlama[15]	13B	<u>90.36</u>	48.96	69.66	14.20	40.71	-	-	-
ChartAst[47]	13B	93.90	65.90	79.90	15.20	41.00	-	<u>91.60</u>	15.50
Vicuna (IC-tuning)	7B	90.32	<u>58.72</u>	<u>74.52</u>	12.43	<u>42.75</u>	97.57	92.77	17.65

Table 2. The evaluation results of traditional chart benchmarks. The top two performances for each benchmark are highlighted in **bold** and underlined, respectively. For comparison with other models, the performance of Chart-to-text, Chart-to-table, OpenCQA is evaluated after finetuning on each train split.

This may be due to the ground truths of the Chart-to-text being very short and covering only a portion of the chart without visual information. However, our model-generated chart summarizations include visual information and provide a comprehensive explanation of the entire chart, which may result in a lower BLEU score.

Furthermore, the BLEU metric have been demonstrated to exhibit a low correlation with human evaluations, particularly in open-ended text generation tasks. In contrast, G-Eval [40] has recently been widely adopted in various studies for evaluating texts generated by LLMs [15, 22, 59]. Hence, we conduct an additional evaluation of chart comprehension using G-Eval.

Chart Summarization and OpenCQA evaluation using G-Eval. For G-Eval, we employ two benchmarks: Chart Summarization and OpenCQA. For the Chart Summarization task, we collect 140 charts from seven diverse sources⁷, instead of using the Chart-to-text dataset, which only includes charts from Pew and Statista. Meanwhile, we use OpenCQA, a benchmark already utilized in traditional chart evaluations, because it has human-written questions.

For the Chart Summarization task, we evaluate the model

⁷The charts are from Pew, Statista, OECD, OWID, PlotQA, ChartInfo, Data Aug., all of which are used in Unichart.

	Chart Summarization	OpenCQA
ChartLlama[15]	2.54	2.56
ChartAst[47]	3.28	3.81
Vicuna (IC-tuning)	3.85	4.11
- w/o unichart	3.14	3.54
- w/o stage 2	<u>3.82</u>	<u>4.01</u>
- w/o stage 3	3.41	3.65

Table 3. The evaluation result of Chart Summarization and OpenCQA using G-Eval.

on the criteria of factual correctness, informativeness, and fluency, while for the OpenCQA task, we assess it based on the criteria of factual correctness, relevance, and fluency.⁸ We use a 5-point scale ranging from 1 to 5 for these criteria. Table 3 shows the average values of three criteria for each task, obtained by performing G-Eval five times.⁹

We compare *Vicuna (IC-tuning)* with other LLMs tuned with ChartLlama and ChartAst, which shows great performance on traditional chart benchmarks.¹⁰ *Vicuna (IC-*

⁸A detailed description of the evaluation criteria is in the Appendix I.

⁹The scores for each criteria are shown in Table 14.

¹⁰ChartInstruct is also tuned LLM with chart instruction-following data. However, we were unable to reproduce this model because it was not publicly available.

tuning) shows the highest scores for both Chart Summarization and OpenCQA tasks, with scores of 3.85 and 4.11, respectively. This demonstrates that, unlike in traditional chart benchmarks, our IC-tuning method is highly effective not only in OpenCQA but also in Chart Summarization task.

Additionally, we conduct several experiments to validate the effectiveness of the IC-tuning model architecture and its training strategy. First, *w/o unichart*, where we use the CLIP encoder instead of Unichart encoder, shows a significant drop in performance.¹¹ This suggests that the Unichart encoder, pre-trained on diverse chart data, is more suitable for chart understanding than CLIP encoder. Furthermore, *w/o stage 2* and *w/o stage 3*, where we omit each respective training stage, also show a decrease in performance. Notably, *w/o stage 3* shows a substantial performance drop, indicating that visual-aware chart instruction following data is crucial for chart comprehension performance.

Ablation studies on chart, image, and text modality comprehension. To verify the effectiveness of IC-tuning for comprehending chart, image, and text modalities, we conduct ablation studies across various training scenarios using chart, vision, and language benchmarks, as shown in Table 4. For the chart benchmark, we reuse the traditional chart benchmark mentioned earlier, and the details of the vision and language benchmarks can be found in Appendix F.

In the continual learning scenario, we further train LLaVA 7B with the chart instruction-following data of stage 2 to enhance its comprehension of the chart modality. For this, *CLIP Post-tuning* uses a CLIP encoder for the chart modality. In contrast, *Unichart tuning* utilizes a Unichart encoder for the chart modality instead of the CLIP encoder. In the end-to-end training scenario, we train Vicuna 7B with a mixture of multimodal instruction-following data of stage 2 to enhance its comprehension of the image and chart modalities. For this, *CLIP(image, chart) tuning* uses a single CLIP encoder to connect both chart and image modalities to Vicuna. In contrast, *CLIP(image), Unichart(chart) tuning* employs a CLIP encoder for the image modality and a Unichart encoder for the chart modality. Note that we fine-tune LLM and projection layers in both scenarios.

In the continual learning scenario, the overall chart benchmark performance of *Unichart tuning* surpasses that of *CLIP Post-tuning*. Additionally, the performance degradation from LLaVA is smaller with *Unichart tuning* compared to *CLIP Post-tuning* on vision and language benchmarks, suggesting that *Unichart tuning* experiences less catastrophic forgetting during the continual learning of LLaVA. In an end-to-end training scenario, *CLIP(image), Unichart(chart) tuning* outperforms *CLIP(image, chart) tuning* approach across all chart, vi-

¹¹We use the CLIP encoder from LLaVA-NeXT with the AnyRes technique [29].

sion, and language benchmarks. This represents that *CLIP(image), Unichart(chart) tuning* is less affected by task interference across different modalities during multi-task tuning on a mixture of multi-modal instruction following data. Consequently, utilizing CLIP encoder for the image modality and the Unichart encoder for the chart modality separately proves to be an effective model architecture for maximizing performance across all modalities in both scenarios.

4.3. Question answering over multi-modal contexts evaluation

We evaluate our models on the test split of ICT-QA dataset to assess its performance on QA tasks over multi-modal contexts. We use GPT-4o to assess factual correctness, relevance, and fluency.¹²

The overall evaluation results are shown in Table 5.¹³ We compare *Vicuna (IC-tuning)*, *Vicuna (IC-tuning + ICT-QA)*, which is a model post-tuned with ICT-QA dataset after IC-tuning, and LLaVA-NeXT-Interleave [30]. LLaVA-NeXT-Interleave, fine-tuned on the M4-Instruct dataset, achieves state-of-the-art performance on multi-image tasks. Since treating charts as images makes ICT-QA similar to a multi-image task, we select LLaVA-NeXT-Interleave as a comparison model.¹⁴

The Chart QA score of *Vicuna (IC-tuning)* is 3.93, compared to 3.75 for LLaVA-NeXT-Interleave. This suggests that *Vicuna (IC-tuning)* benefits from its strong chart comprehension, thanks to IC-tuning. On the other hand, *Vicuna (IC-tuning)* scores 3.34 in Image QA, lower than LLaVA-NeXT-Interleave’s 3.71. LLaVA-NeXT [38], the base model of LLaVA-NeXT-Interleave, uses higher resolution images with the AnyRes technique and is trained with larger visual instruction-following data compared to IC-tuning, possibly explaining why *Vicuna (IC-tuning)* scores lower in Image QA. To compare the Image QA performance of LLaVA with *Vicuna (IC-tuning)*, considering that LLaVA cannot process images and charts simultaneously, we exclude the chart and use only the image and text as input. In this configuration, LLaVA scores 3.92, and *Vicuna (IC-tuning)* scores 3.97. These very similar scores indicate that the IC-tuning method helps maintain Image QA performance at a level comparable to LLaVA.

The average G-Eval score for *Vicuna (IC-tuning + ICT-QA)* across all modalities is 4.30, which is significantly higher than the 3.77 score of LLaVA-NeXT-Interleave. This demonstrates that the ICT-QA dataset is highly effective in improving the performance of all modality QAs within

¹²The prompts used for this evaluation can be found in Appendix L.3.

¹³The scores for each criterion on ICT-QA are shown in Table 15 of the Appendix K.

¹⁴We intended to include ChartLlama and ChartAst in our evaluation, but did not use them; they sometimes did not follow the instructions of ICT-QA.

	ChartQA (<i>RA</i>)			Chart-to-text (<i>BLEU</i>)		Chart-to-table (<i>RNSS</i> <i>RMS_{F1}</i>)		OpenCQA (<i>BLEU</i>)	Vision Benchmarks	Language Benchmarks
	aug.	human	avg.	Pew	Statista	ChartQA		OpenCQA	Avg.	Avg.
Continual learning										
LLaVA	13.92	18.08	16.00	11.23	41.65	68.16	1.18	4.63	63.41	52.27
- <i>CLIP post-tuning</i>	61.36	36.96	49.16	11.16	41.50	91.11	43.10	13.41	59.87	48.64
- <i>Unichart tuning</i>	<u>92.72</u>	<u>46.96</u>	<u>69.84</u>	12.94	43.34	<u>97.11</u>	<u>91.62</u>	14.23	62.22	50.23
End-to-end										
Vicuna	-	-	-	-	-	-	-	-	-	53.68
- <i>CLIP(image, chart) tuning</i>	63.04	34.56	48.80	12.20	41.53	92.08	44.44	13.17	61.88	53.02
- <i>CLIP(image), Unichart(chart) tuning</i>	92.80	47.68	70.24	<u>12.70</u>	<u>43.24</u>	97.32	91.64	<u>14.14</u>	<u>62.96</u>	<u>53.25</u>

Table 4. **The ablation studies on IC-tuning for Chart, Vision and Language benchmarks.** For Chart-to-text, the performance is evaluated after finetuning on train split.

Model	Image QA	Image QA (w/o chart)	Chart QA	Text QA	Avg.
LLaVA (Vicuna 7B)	-	3.92	-	-	-
LLaVA-NeXT-Interleave (Qwen 7B)	3.71	-	3.75	3.84	3.77
Vicuna (IC-tuning)	3.34	3.97	3.93	4.02	3.77
Vicuna (IC-tuning + ICT-QA)	4.17	-	4.23	4.51	4.30
- <i>w/o sp token, interleave</i>	4.03	-	4.12	4.40	4.18

Table 5. **The evaluation results on ICT-QA.** The G-Eval scores are presented for each modality QA within a multi-modal context. In the case of Image QA (w/o chart), it represents the score of Image QA when only the image and text are provided as input, excluding the chart.

	Image QA			Chart QA			Text QA		
	Relevant	Full	Difference	Relevant	Full	Difference	Relevant	Full	Difference
Single-modal QA									
Vicuna (IC-tuning)	3.81	3.31	0.50	4.05	4.04	0.01	4.26	4.02	0.24
Vicuna (IC-tuning + ICT-QA)	3.89	3.80	0.09	3.94	4.06	-0.12	4.51	4.51	0.00
Cross-modal QA									
Vicuna (IC-tuning)	4.19	3.36	0.84	3.94	3.71	0.22			
Vicuna (IC-tuning + ICT-QA)	4.43	4.35	0.08	4.54	4.57	-0.04			

Table 6. **The ablation studies on ICT-QA for single-modal and cross-modal QA.** “Relevant” denotes the performance when only the relevant context is provided as input. “FULL” denotes the performance when full context is provided as input. For example, in the case of image QA for single-modal QA, “Relevant” indicates when only image is given, while “FULL” indicates when the image, chart, and full text are all given. “Difference” denotes the performance difference between “Relevant” and “FULL”. A smaller “Difference” indicates that the model has a stronger ability to filter out irrelevant details and focus on the essential information.

multi-modal contexts. Additionally, the *w/o sp token, interleave* model, where no special tokens to distinguish modalities are used and modality embeddings are placed at the front, shows an average performance of 4.18, slightly lower than when these techniques are applied. This suggests that these techniques positively impact performance.

In Table 6, we conduct ablation studies to analyze the performance improvement attributed to ICT-QA for single-modal and cross-modal questions. We compare the G-Eval scores when only the relevant context, specifically the part containing the answer to the question, is provided versus when the full context is given.

Since the full context includes irrelevant information, a smaller score difference between when the full context is

given and the relevant context is given indicates that the model is more robust in identifying and focusing on the necessary parts of the multi-modal context. These differences for *Vicuna (IC-tuning + ICT-QA)* are significantly smaller than those for *Vicuna (IC-tuning)* across all modality QAs, including both single-modal and cross-modal questions. This suggests that the ICT-QA dataset enhances the model’s robustness to multi-modal contexts that include irrelevant information.

Furthermore, when the full context is provided, the scores of *Vicuna (IC-tuning + ICT-QA)* across all modality QAs, including both single-modal and cross-modal QA, are higher than those of *Vicuna (IC-tuning)*. This indicates that the ICT-QA dataset contributes to the performance im-

provement for both single-modal and cross-modal questions. Notably, in cross-modal QA, the scores of *Vicuna (IC-tuning + ICT-QA)* in image QA and chart QA, at 4.35 and 4.57 respectively, show a particularly large performance gap compared to *Vicuna (IC-tuning)*, which scores 3.36 and 3.71. This highlights that the ICT-QA dataset is highly effective for cross-modal questions.

Error analysis. To identify the challenging aspects our model faced, we analyze the results of *Vicuna (IC-tuning + ICT-QA)*. The challenging examples, E1, E2, E3, and E4, are shown in Figure 12 of Appendix C.

- **Color recognition within charts** Our model generally performs well in color recognition as shown in D2 of the Figure 3, which is provided in Appendix B.1, but occasionally incorrectly recognizes colors within charts. As shown in E1, while our model accurately recognizes the value corresponding to the question, it provides an incorrect color for the line. This issue arises because ICT-QA is generated by GPT-4V, which sometimes fails to accurately interpret graphs with diverse colors or line styles.¹⁵
- **Uncommon types of charts** Our model struggles to accurately understand uncommon types of charts. In the case of E2, the chart type differs slightly from the chart types commonly used in the IC-tuning and ICT-QA datasets. Although our model understands the intent of the question, it provides incorrect values. Enhancing ICT-QA to include a broader range of chart types could improve performance in such cases.
- **Factual errors** Our model shows significant performance improvements, but it still occasionally generates factually incorrect statements. In the case of E3, while our model accurately identifies the values intended in the question, it generate incorrect comparison between them. Upon analyzing the cause of this issue, we discovered that certain chart-text pairs from the Unichart data [45] used terms like “low”, “high”, “increase”, and “decrease” with reversed meanings. By filtering out such erroneous data, these issues could potentially be addressed.

5. Limitations and Opportunities

- **The quality of the chart instruction tuning data** In IC-tuning, we trained the model with a mixture of existing chart instruction-following data in stage 2 and with our visual-aware chart instruction-following data in stage 3. A significant portion of these data was generated by LLMs. Upon close examination, we found that while most of the data is of high quality, some instances occasionally contain errors. This is because LLMs can sometimes produce hallucinations when generating chart

instruction-following data. We believe that performance can be improved by filtering out these erroneous data.

- **Long sequence length of modality tokens** For an image and a chart, CLIP and Unichart encoders generates 576 and 900 tokens, respectively. Therefore, even if we only use one image and one chart as input, the sequence length reaches 1476 tokens. This could become a bigger issue in future research if we use multiple images and charts as inputs. Therefore, it would be beneficial to explore methods to reduce the token length during modality projection.
- **The restricted styles of charts** In IC-tuning, the chart instruction-following data was collected from various sources, including existing datasets such as Unichart and ChartQA. The ICT-QA dataset, based on articles only from the Pew Research Center, also includes various types of charts (e.g., bar, line, pie). However, compared to IC-tuning, it utilizes a more limited range of data sources, resulting in relatively fewer chart styles. Building the ICT-QA dataset based on a broader range of data sources would be beneficial.
- **The gap between the multi-modal contexts and real-world scenarios** In ICT-QA, the multi-modal contexts includes up to one image and one chart, and does not accommodate multiple images or charts. However, in real-world scenarios, documents often contain multiple images and charts. Additionally, when training the ICT-QA model, the position of images and charts is not considered. Future research should focus on incorporating multiple images and charts as well as their positions.

6. Conclusion

In this paper, we propose the IC-tuning method, which includes the proposal of visual-aware chart instruction-following data and a model architecture utilizing image and chart-specific encoders. Through experiments, we verify that this approach improves chart comprehension performance while minimizing the impact on image and language comprehension. Furthermore, we propose the ICT-QA dataset and show that tuning with this data improves the quality of answers for single-modal and cross-modal questions from multi-modal contexts. In future research, we will focus on bringing our work closer to real-world applications. For example, we plan to expand multi-modal contexts to include multiple images and charts within much longer documents.

References

- [1] Sharegpt. <https://sharegpt.com/>, 2023. 4, 21
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances*

¹⁵<https://platform.openai.com/docs/guides/vision/limitations>

- in *neural information processing systems*, 35:23716–23736, 2022. 12
- [3] 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. 1, 12, 18
- [4] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16495–16504, 2022. 12
- [5] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 1, 12
- [6] Wenhui Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Wang. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. *arXiv preprint arXiv:2004.07347*, 2020. 12
- [7] Wenhui Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks, 2023. 4
- [8] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 4, 12
- [9] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. 18
- [10] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 18
- [11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 12
- [12] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2024. 12, 18
- [13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 12, 18, 21
- [14] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3608–3617, 2018. 18
- [15] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation, 2023. 1, 5, 12, 18
- [16] Darryl Hannan, Akshay Jain, and Mohit Bansal. Many-modalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7879–7886, 2020. 2, 12
- [17] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 18
- [18] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 18, 21
- [19] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 12
- [20] Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. Opencqa: Open-ended question answering with charts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11817–11837, 2022. 2, 4, 17, 21
- [21] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization, 2022. 4, 17, 21
- [22] Tatsuki Kawamoto, Takuma Suzuki, Ko Miyama, Takumi Meguro, and Tomohiro Takagi. Application of frozen large-scale models to multimodal task-oriented dialogue, 2023. 5
- [23] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 21
- [24] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoon Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer, 2022. 3
- [25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 21

- [26] Kenton Lee, Mandar Joshi, Iulia Turc, Hexiang Hu, Fangyu Liu, Julian Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding, 2023. 5, 18
- [27] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 12
- [28] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 18
- [29] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, 2024. 6, 12
- [30] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next: Tackling multi-image, video, and 3d in large multimodal models, 2024. 2, 6, 12
- [31] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 12
- [32] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 18
- [33] Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021. 18
- [34] Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhui Chen, Nigel Collier, and Yasemin Altun. Deplot: One-shot visual language reasoning by plot-to-table translation, 2023. 3, 4, 18, 19
- [35] Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12756–12770, Toronto, Canada, 2023. Association for Computational Linguistics. 5, 12, 18
- [36] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*, 2023. 1, 4, 12, 18, 21
- [37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 12, 17, 18, 21
- [38] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 6
- [39] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 18
- [40] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, 2023. 2, 5
- [41] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 18
- [42] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 21
- [43] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 21
- [44] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. 2, 4, 17, 18, 21
- [45] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14662–14684, Singapore, 2023. Association for Computational Linguistics. 1, 2, 3, 5, 8, 12, 17, 18, 21
- [46] Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. Chartinstruct: Instruction tuning for chart comprehension and reasoning, 2024. 5, 12, 18
- [47] Fanqing Meng, Wenqi Shao, Quanfeng Lu, Peng Gao, Kaipeng Zhang, Yu Qiao, and Ping Luo. Chartassistant: A universal chart multimodal language model via chart-to-table pre-training and multitask instruction tuning. *arXiv preprint arXiv:2401.02384*, 2024. 1, 4, 5, 12, 18
- [48] Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2020. 21
- [49] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 international conference on document analysis and recognition (ICDAR)*, pages 947–952. IEEE, 2019. 21
- [50] OpenAI. Gpt-4 technical report, 2023. 1, 2
- [51] Matt Post. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, 2018. Association for Computational Linguistics. 4, 17

- [52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. [1](#), [3](#), [12](#)
- [53] Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021. [18](#)
- [54] Kate Sanders, David Etter, Reno Kriz, and Benjamin Van Durme. Multivent: Multilingual videos of events and aligned natural text. *Advances in Neural Information Processing Systems*, 36, 2024. [12](#)
- [55] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. [21](#)
- [56] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer, 2020. [21](#)
- [57] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. [18](#)
- [58] Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. MIMOQA: Multimodal input multimodal output question answering. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 5317–5332, 2021. [12](#)
- [59] Jae-hee So, Joonhwan Chang, Eunji Kim, Junho Na, JiYeon Choi, Jy-yong Sohn, Byung-Hoon Kim, and Sang Hui Chu. Aligning large language models for enhancing psychiatric interviews through symptom delineation and summarization. *arXiv preprint arXiv:2403.17428*, 2024. [5](#)
- [60] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021. [2](#), [12](#)
- [61] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *corr, abs/2302.13971*, 2023. doi: 10.48550. *arXiv preprint arXiv:2302.13971*, 2023. [12](#)
- [62] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrubti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. [4](#)
- [63] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shrubti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. [12](#)
- [64] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl: Modularization empowers large language models with multimodality, 2024. [1](#), [12](#)
- [65] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. [18](#)
- [66] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019. [18](#)
- [67] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. [1](#), [12](#)