

# MMAT-1M: A Large Reasoning Dataset for Multimodal Agent Tuning

## Supplementary Material

### A. Dataset Statistics

Table 1 presents detailed statistics of the datasets utilized to construct MMAT-1M. These datasets are curated from various prominent sources, including Visual CoT [28], LLaVA-CoT [35], The Cauldron [14], TabMWP [20], and InfoSeek [4], collectively contributing to an extensive multimodal reasoning dataset. Specifically, the table enumerates the composition, the number of data entries, and the corresponding QA pairs of each sub-dataset. In total, the MMAT-1M dataset comprises 1,090,263 QA pairs, indicating substantial coverage and diversity in multimodal tasks.

Category	Number of Data Entries	Number of QA Pairs
Visual CoT [28]		
Birds-200-2011 [33]	10.1k	10.1k
DocVQA [22]	33.5k	33.5k
DUDE [32]	11.7k	11.7k
Flickr30K [27]	135.7k	135.7k
GQA [8]	98.1k	98.1k
InfographicsVQA [23]	15.1k	15.1k
Open images [13]	43.1k	43.1k
SROIE [7]	2.5k	2.5k
TextCap [29]	32.2k	32.2k
TextVQA [30]	18.5k	18.5k
Visual7W [40]	30.5k	30.5k
VSR [16]	3.4k	3.4k
LLaVA-CoT [35]		
ShareGPT4V [3]	31.3K	67.9k
ChartQA [21]	17.0k	25.6k
A-OKVQA [21]	16.1K	99.9K
AI2D [10]	11.4k	11.4k
GeoQA+ [2]	11.4k	11.4k
ScienceQA [19]	5.6k	5.6k
DocVQA [22]	4.0k	31.3k
PISC [15]	0.9k	0.9k
CLEVR [9]	0.5k	0.5k
CLEVR-Math [6]	0.5k	0.5k
The Cauldron [14]		
HatefulMemes [12]	8.5k	8.5k
Screen2Words [34]	15.7k	15.7k
ST-VQA [1]	17.2k	23.1k
VisText [31]	10.0k	10.0k
WikiSQL [39]	75.0k	86.2k
WTQ [26]	38.2k	44.1k
IconQA [18]	27.3k	29.8k
RAVEN [37]	20.9k	20.9k
Inter-GPS [17]	1.3k	1.8k
TQA [11]	1.5k	6.5k
TabMWP [20]		
TabMWP [20]	23.1k	23.1k
InfoSeek [4]		
InfoSeek [4]	131.4k	131.4k

Table 1. Detailed statistics of datasets included in MMAT-1M.

### B. Data Engine

This section introduces the API operators, including scene graph-based image caption generation. It then details the prompts and their designs underlying the data engine, clearly distinguishing between rationale generation and reflection components.

Inspired by the CCoT [25], GPT-4o is leveraged to construct a scene graph and derive an image caption, enhancing semantic understanding and compositional reasoning. The scene graph prompt and image caption prompt are depicted in 1.

#### Scene Graph Prompt

{QUESTION}

For the provided image and its associated question, generate only a scene graph in JSON format that includes the following:

1. Objects that are relevant to answering the question
2. Object attributes that are relevant to answering the question
3. Object relationships that are relevant to answering the question

Example of a valid JSON response:

```
{
  "objects": ["object_name1", "object_name2", ...],
  "attributes": [
    {"name": "object_name", "value": "attribute_value"},
    ...
  ],
  "relationships": [
    {"object_name1": {"relation_type": ["object_name2"]}},
    ...
  ]
}
```

STRICTLY FOLLOW THE JSON RESPONSE FORMAT.

#### Image Caption Prompt

Scene Graph: {SCENE\_GRAPH}

Use the scene graph as context to write a concise description for this image.

Figure 1. Scene graph and caption generation prompts.

For rationale generation, we guide GPT-4o through a structured and adaptive multi-stage reasoning process. During inference, the model dynamically invokes multimodal operators, including Image Caption, Open-Vocabulary Ob-

jection Detection (OVD), Optical Character Recognition (OCR), and Face Detection, while also leveraging retrieval-augmented generation when needed. Each reasoning step is explicitly documented in structured JSON format to maintain transparency and logical coherence. Figure 2 illustrates the comprehensive design of the rationale generation prompt.

**Data Generation Prompt**

You are an advanced multimodal reasoner. Your task is to reason step by step to answer the user's question using the available operators and rags, finally get the standard answer logically.

In each step, decide which operators to call for the next step. You may choose none, one, or multiple operators. The operator results will be provided as input in the following step.

If all information provided by the operators are useless, then use RAG (Retrieval Augmented Generation) to solve the problem. It should be noted that rag only supports text-to-text search, not text-to-image or image-to-image search.

Respond in JSON format with the following fields:

- thought: A concise reasoning step with reflection on previous results.
- operators: An array of operators to use next (e.g., "image\_caption", "label\_bbox", "ocr", "face\_detection"). Leave it empty if no operators are needed.
- rag: An dict of RAG queries that can help resolve the issue. (e.g. {"needed": true, "query": "A query can help reasoning and solve the promblem."})
- next\_action: Either "continue" if further steps are needed, or "final\_answer" if you have the complete solution.

Example of a valid JSON response:

```
{
  "thought": "Identifying key information and reflecting on prior results",
  "operators": ["image_caption"],
  "rag": {"needed": false, "query": null},
  "next_action": "continue"
}
```

Notice: you are merely using reasoning to approach the standard answer, but you should not use the standard answer directly or use the words like standard answer. The thought should be concise and logical.

Figure 2. Rationale generation prompt.

For reflection, we designed two targeted prompts to enhance reasoning robustness. The general reflection prompt is intended to detect and correct reasoning cheating behaviors. Specifically, it prompts GPT-4o to critically exam-

ine cases where its reasoning process artificially aligns with given answers rather than deriving them through genuine inference. The prompt explicitly requests that the model identify and articulate any logical inconsistencies in its reasoning. The detailed structure of the general reflection prompt is provided in Figure 3.

**General Reflection Prompt**

You are an image analysis and reasoning system. Based on the input image and question, you first determine the approach to solve the problem, then invoke external tools to assist in providing an answer, and ultimately combine the analysis process to give your response. The entire reasoning process takes the form of a multi-round dialogue, with the results as follows:

```
{CONVERSATIONS_WO_ANSWER}
```

As a user, I have noticed that some of your reasoning processes may involve cheating.

A typical manifestation is: you think and invoke external tools, but due to the inadequacy of the external tools or the limitations of your own model's capabilities, you actually cannot obtain the correct answer. However, in order to appear as if you can derive the correct answer, you pretend in your final reasoning conclusion that you have derived the correct answer. This constitutes cheating.

I hope you can review the above answers to determine if such cheating behavior exists. If it does, please provide the reason.

Output format:

```
{
  "reason": "The reason for the existence of cheating behavior"
}
```

The content of "reason" is is why you think the above reasoning process involves logical cheating.

If there is no cheating behavior, the output will be empty: {}.

Figure 3. General reflection prompt.

The math reflection prompt specifically targets completeness issues in mathematical reasoning. It instructs GPT-4o to carefully inspect its mathematical derivations, identifying instances where crucial calculation steps might be omitted. By prompting the model to supplement missing derivations explicitly, this ensures the integrity and clarity of mathematical reasoning. Figure 4 depicts the math reflection prompt's structure in detail.

Our MMAT-1M dataset supports both one-turn and multi-turn reasoning frameworks, each with a dedicated system prompt.

### Math Reflection Prompt

You are an image analysis and mathematical reasoning system that provides answers based on input images and questions in the form of multiple rounds of dialogue:

{CONVERSATIONS}.

However, as a user, I'm not good at mathematical derivations. If you omit necessary steps in your reasoning process, I may make mistakes in my reasoning and fail to obtain the final answer.

Especially, I've noticed that in your answers, from the second-to-last round of dialogue:

{CONVERSATION\_Q},

to the last round:

{CONVERSATION\_V},

sometimes the specific mathematical calculation process is omitted.

If this is the case, could you please provide additional reasoning to help me understand? If you believe the above answer is already very detailed and complete, then there is no need to add anything.

Now, please respond in the following format:

```
{
  "reason": "Specific reason",
  "thought": "Supplementary reasoning process/null"
}
```

Where "reason" is the reason why additional reasoning is needed or not needed, and "thought" is the supplementary reasoning process. If no additional reasoning is needed, it must be "null".

Note:

1. the output format must be a json.
2. You need to carefully examine whether the logic from the second-to-last step to the final step is sufficient, there is no need to supplement it with additional reasoning for each question.

Figure 4. Math reflection prompt.

The one-turn rationale and reflection (ORR) prompt enables the model to complete the entire reasoning process within a single inference step. This design ensures efficiency while maintaining strong reasoning capabilities. The structured output format ensures clarity and consistency. The specific prompt format for ORR is shown in Figure 5.

The rationale and reflection (RR) prompt guides the model through an iterative reasoning process, dynamically selecting multimodal operators and retrieving external knowledge when necessary. This approach enhances interpretability and reasoning depth. Figure 6 outlines the specific format designed for RR.

To assess potential GPT-4o hallucinations introduced

### ORR System Prompt

You are an advanced multimodal reasoner. Your task is to reason step by step to answer the user's question using the available operators, finally get the answer logically.

Respond in the format like this:

<THOUGHT>your reasoning</THOUGHT>

<ANSWER>The answer is: your answer</ANSWER>

Figure 5. One-turn rationale and reflection (ORR) prompt.

### RR System Prompt

You are an advanced multimodal reasoner. Your task is to reason step by step to answer the user's question using the available operators and rags, finally get the answer logically.

In each step, decide which operators to call for the next step. You may choose none, one, or multiple operators. The operator results will be provided as input in the following step.

You can try to use RAG (Retrieval Augmented Generation) to search more useful information.

During the reasoning steps, your output format should be as follows:

<THOUGHT>reasoning</THOUGHT>

<OPERATOR>names of the operators</OPERATOR>

<RAG>query for rag</RAG>

<SIGNAL>signal of the next action</SIGNAL>

Note, if there is no need to call operators or RAG, <OPERATOR> or <RAG> can be omitted from the output.

In the final step, your response must be the ultimate answer to the initial question.

Figure 6. Rationale and reflection (RR) prompt.

during the rationale and reflection generation stages, we performed a large-scale quality evaluation of the entire MMAT-1M dataset. Specifically, we employed the Doubao-1.5-Vision-Pro-32K model to evaluate all samples against five criteria: coherence, relevance, accuracy, completeness, and image-text integration. Results indicate that over 89% of the samples exhibit high-quality reasoning. Figure 7 presents the prompt used in this evaluation.

## C. Training Hyperparameters

In this section, we present the main training parameters for multiple models. For all models, including Llama-3.2-11B-Vision-Instruct [24], MiniCPM-V-2.6 [36], and the InternVL2.5 series [5], we adopt the same training configura-

### Reasoning evaluation prompt

You are an experienced evaluator of multimodal reasoning processes. Your task is to assess the quality of a structured multimodal Chain-of-Thought (CoT) reasoning process and assign an overall quality score.

Evaluation Criteria:

Coherence: Logical and fluent step-by-step reasoning.

Relevance: All steps directly contribute to answering the query.

Accuracy: Reasoning leads to the correct final answer.

Completeness: Proper use of multimodal operators (e.g., Image Caption, OVD, OCR, Face Detection) and Retrieval-Augmented Generation (RAG).

Visual Integration: Effective fusion of image and text-based reasoning.

Assign a score from 0.0 (poor) to 1.0 (excellent), accurate to one decimal place.

Output Format:

Provide your evaluation in JSON format:

```
{
  "quality_score": "A numeric score between 0.0 and 1.0,
rounded to one decimal place.",
  "scoring_reason": "A brief explanation for the score."
}
```

Figure 7. Reasoning evaluation prompt.

tion and use the open-source framework ms-swift [38] for training. The specific parameters are shown in Table 2.

Parameter	Value
train_type	LoRA
num_train_epochs	1
train_batch_size	1
gradient_accumulation_steps	1
learning_rate	$4 \times 10^{-5}$
weight_decay	0.1
max_length	16384
torch_dtype	BF16
seed	42
deepspeed	ZeRO-2

Table 2. Configuration of hyperparameters used in training.

## D. More Qualitative Examples

In this section, we present additional qualitative examples highlighting the practical benefits of iterative reasoning, self-reflection, and one-turn reasoning.

Figure 8 illustrates how the model refines its reasoning to correct an OCR error. The initial OCR result misidentifies the text as “ADEDNI”, but by leveraging contextual understanding and common knowledge, the model correctly recognizes it as “CALIFORNIA”. This highlights the model’s ability to detect and correct errors through iterative reasoning.

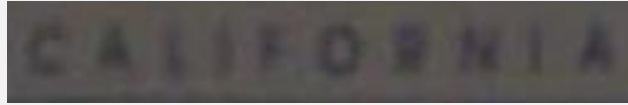
Figure 9 showcases how reflection improves mathematical reasoning. Initially, the model applies a direct calculation to determine the area of a parallelogram, but realizes its approach lacks mathematical rigor. Through self-reflection, it revises its reasoning and correctly applies the sine function, leading to an accurate computation of the area.

Figure 10 illustrates how the model integrates external signals such as image caption and object detection to improve decision-making. In a traffic scene, the baseline model incorrectly determines the direction of the closest car. However, by analyzing additional image information, the model correctly identifies that the car is approaching from the opposite lane, leading to the correct conclusion.

These examples underscore the model’s strengths in error correction, reasoning refinement, and effective use of external knowledge for improved decision-making.

## E. Limitations and Future Work

Despite its strengths, MMAT-1M has certain limitations. The reliance on high-quality rationale training data and a fixed set of tool usage may restrict its generalization ability. Additionally, while the dataset demonstrates strong performance on current benchmarks, further testing is needed to evaluate its adaptability to more MLLMs and more complex scenarios. Nevertheless, MMAT-1M sets a new pipeline for multimodal agent tuning, and its performance on challenging tasks underscores its potential as a foundational visual training dataset for future research.



**Question:**

what is written in the image?.

**InternVL-2.5-8B-RR:**

**Q:**what is written in the image?

Answer the question using a single word or phrase. Please start your reasoning.

**A:**<THOUGHT>To determine the text in the image, using OCR will be the most effective approach.</THOUGHT>

<OPERATOR>ocr</OPERATOR>

<SIGNAL>continue</SIGNAL>

**Q:**Here are ocr results from the previous step:



- ocr information: **ADEDNI**

Please continue your reasoning.



**A:**<THOUGHT>**The OCR result 'ADEDNI' seems to be a misinterpretation.**

**The text might be 'CALIFORNIA' based on the context and common knowledge.**</THOUGHT>

<SIGNAL>end</SIGNAL>

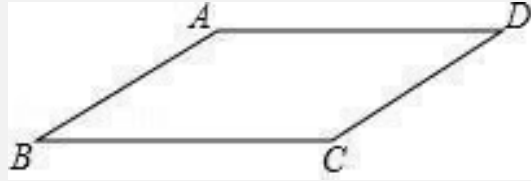
**Q:**Based on the above information, please summarize and give the answer to the initial question: what is written in the image?

Answer the question using a single word or phrase.



**A:****california**

Figure 8. Example of iterative rationale result. The OCR recognition is erroneous, but the rationale process corrects the mistake.



**Question:**

Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.  
 Question: As shown in the figure, in the parallelogram ABCD, it is known that  $AB = 6.0$ ,  $BC = 9.0$ ,  $\angle B = 30.0$ , then the area of the parallelogram ABCD is ()  
 A. 12  
 B. 18  
 C. 27  
 D. 54  
 Answer with the option's letter from the given choices directly.

**InternVL-2.5-8B-RR:**

Q: Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.  
 Question: As shown in the figure, in the parallelogram ABCD, it is known that  $AB = 6.0$ ,  $BC = 9.0$ ,  $\angle B = 30.0$ , then the area of the parallelogram ABCD is ()  
 A. 12  
 B. 18  
 C. 27  
 D. 54  
 Answer with the option's letter from the given choices directly. Please start your reasoning.

A: <THOUGHT>The image shows a parallelogram ABCD with given side lengths and angle. To find the area, we can use the formula:  $\text{Area} = \text{base} \times \text{height}$ . The base is AB, and the height can be found using the sine of angle B.  
 B: <THOUGHT>  
 <SIGNAL>continue</SIGNAL>

Q: No additional operator or rag results were selected in the previous step. Please continue your reasoning.



A: <THOUGHT>Wait a bit, the way I was thinking about it might not have enough math backing it up. I'll improve it.  
 <THOUGHT>  
 <SIGNAL>reflection</SIGNAL>

Q: No additional operator or rag results were selected in the previous step. Please make a reflection.



A: <THOUGHT>To calculate the area of the parallelogram ABCD, we use the formula:  $\text{Area} = \text{base} \times \text{height}$ . Here, the base is  $AB = 6.0$ . The height can be found using the sine of angle B, which is  $30^\circ$ . The height is  $BC \times \sin(30^\circ) = 9.0 \times 0.5 = 4.5$ . Therefore, the area of the parallelogram is  $6.0 \times 4.5 = 27$ . This confirms the answer is C.  
 <THOUGHT>  
 <SIGNAL>end</SIGNAL>

Q: Based on the above information, please summarize and give the answer to the initial question: Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.  
 Question: As shown in the figure, in the parallelogram ABCD, it is known that  $AB = 6.0$ ,  $BC = 9.0$ ,  $\angle B = 30.0$ , then the area of the parallelogram ABCD is ()  
 A. 12  
 B. 18  
 C. 27  
 D. 54  
 Answer with the option's letter from the given choices directly.



A: C

Figure 9. Example of reflection result. The inclusion of the reflection makes the mathematical derivation more complete and accurate.



**Question:**

Is the car closest to us driving in the same direction as us or in the opposition direction from us.

- A. Same direction.
- B. Opposite direction.

Answer with the option's letter from the given choices directly.

**InternVL-2.5-8B-Baseline:**

**Q:**Is the car closest to us driving in the same direction as us or in the opposition direction from us.

- A. Same direction.
- B. Opposite direction.

Answer with the option's letter from the given choices directly.

**X** A: A

**InternVL-2.5-8B-ORR**

**Q:**Is the car closest to us driving in the same direction as us or in the opposition direction from us.

- A. Same direction.
- B. Opposite direction.

Answer with the option's letter from the given choices directly.

Use the following API outputs to help answer the question:

<|begin\_api\_invoke>image\_caption<|end\_api\_invoke>

<|begin\_result\_of\_api>

A divided road features vehicles traveling in opposite directions, with the closest car approaching from the opposite lane.

<|end\_result\_of\_api>

<|begin\_api\_invoke>label\_bbox<|end\_api\_invoke>

<|begin\_result\_of\_api>

{'closest\_car': [[0.0, 0.46, 0.2, 0.72]], 'road\_lane': [[0.17, 0.47, 1.0, 0.99]]}

<|end\_result\_of\_api>

A: <THOUGHT>To determine the direction of the car closest to us, I need to analyze the image for the orientation of the vehicles. **The image caption indicates that the closest car is approaching from the opposite lane, confirming it is driving in the opposite direction.**</THOUGHT>

<ANSWER>The answer is: B</ANSWER>



Figure 10. Example of one-turn rationale result. It correctly answers the question with the aid of image caption information.

## References

- [1] Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gomez, Marçal Rusiñol, C.V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4290–4300, 2019. 1
- [2] Jie Cao and Jing Xiao. An augmented benchmark dataset for geometric question answering through dual parallel text encoding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, Gyeongju, Republic of Korea, 2022. International Committee on Computational Linguistics. 1
- [3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision (ECCV)*, 2024. 1
- [4] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? *arXiv preprint arXiv:2302.11713*, 2023. 1
- [5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3
- [6] Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. In *International Joint Conference on Learning and Reasoning, 16th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy 2022)*, Windsor, UK, September 28-30, 2022, pages 155–170. Technical University of Aachen, 2022. 1
- [7] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019. 1
- [8] 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. 1
- [9] 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 (CVPR)*, 2017. 1
- [10] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer, 2016. 1
- [11] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384, 2017. 1
- [12] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems*, pages 2611–2624. Curran Associates, Inc., 2020. 1
- [13] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 1
- [14] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024. 1
- [15] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. People in social context (pisc) dataset, 2017. Data set. 1
- [16] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 2023. 1
- [17] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 2021. 1
- [18] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2021. 1
- [19] Pan Lu, Swaroop Mishra, Tony 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. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022. 1
- [20] Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*, 2022. 1
- [21] Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. 1
- [22] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceed-*

- ings of the *IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021. 1
- [23] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 1
- [24] AI Meta. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. *Meta AI Blog*. Retrieved December, 20:2024, 2024. 3
- [25] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14420–14431, 2024. 1
- [26] Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China, 2015. Association for Computational Linguistics. 1
- [27] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1
- [28] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 1
- [29] 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. 1
- [30] 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. 1
- [31] Benny J. Tang, Angie Boggust, and Arvind Satyanarayan. VisText: A Benchmark for Semantically Rich Chart Captioning. In *The Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023. 1
- [32] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Paweł Joziak, Rafał Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540, 2023. 1
- [33] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 1
- [34] Bryan Wang, Gang Li, Xin Zhou, Zhourong Chen, Tovi Grossman, and Yang Li. Screen2words: Automatic mobile ui summarization with multimodal learning. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, page 498–510, New York, NY, USA, 2021. Association for Computing Machinery. 1
- [35] Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. Llava-cot: Let vision language models reason step-by-step, 2024. 1
- [36] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 3
- [37] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1
- [38] Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, Wenmeng Zhou, and Yingda Chen. Swift:a scalable lightweight infrastructure for fine-tuning, 2024. 4
- [39] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning, 2017. 1
- [40] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. 1