Decomposed Prompting for Vision and Language Arithmetic Reasoning

Anonymous ACL submission

Abstract

Math problems that involve both vision and language pose a challenging multi-modal task that requires the integration of visual information, textual information, and strong numerical reasoning for adequately solving it. While large language models (LLMs) have achieved impressive performance on arithmetic word problems based solely on text, we show that introducing visual data significantly increases the difficulty. Specifically, the compositional task of counting objects following recognition becomes a formidable hurdle for math problem-solving with large vision-language models (LVLMs). The dual demand of recognizing objects and performing arithmetic reasoning poses a significant challenge, hindering LVLMs from excelling in the overarching task. We show that the commonly employed chain-of-thought (CoT) approach for decomposed reasoning, designed for LLMs, proves ineffective when applied to this multimodal task. As an alternative to demonstration-based CoT we propose a novel decomposition prompting approach, explicitly breaking down the task into two stages as follows. The first stage performs object detection and enumeration referenced within the mathematical problem. The second stage leverages the output from stage one to directly address the posed math question. Our results demonstrate that this approach leads to stark performance improvements on established benchmarks for visual and language arithmetic problems. This breaks the chains of CoT, paving the way towards developing new and novel multimodal breakdown approaches.

1 Introduction

Math word problems (MWPs) constitute a category of reasoning tasks demanding translation from real-world scenarios into mathematical equations, applying arithmetic operations, and interpreting the results back into meaningful solutions (Kushman et al., 2014). Similar to other reasoning tasks such as analogy problems (Webb et al., 2023) and common sense reasoning (Bisk et al., 2020), large language models (LLMs, Brown et al., 2020a; Smith et al., 2022; Chowdhery et al., 2023a) have exhibited impressive reasoning capabilities in tackling MWPs (Gaur and Saunshi, 2023).

Vision and language math problems (VLMP) constitute a multi-modal task that builds upon the foundation of MWP (Lu et al., 2023). Unlike MWP, VLMP necessitates the integration of both visual and linguistic modalities to solve the mathematical problem. The inclusion of two modalities requires diverse cognitive abilities, specifically object recognition from vision and quantitative reasoning guided by textual cues. For instance, when posed with the problem "How many figs will remain in the image if Joe eats one?", accurate detection of the number of figs in the image becomes essential.

Similarly to LLMs, LVLMs evolved from being task-specific (fine-tuned), to instruction-tuned general-purpose models. A commonly used approach to guide LLMs to complete complex tasks is through chain-of-thought (CoT, Wei et al., 2022). CoT does not only give examples of the task and answer, but demonstrates to the model step-by-step how to break the task into smaller more manageable tasks, in order to reach the correct answer. This approach has been applied to LVLM question-answering for the ScienceQA benchmark (Lu et al., 2022a; Zhang et al., 2023).

In this work, we identify a key challenge with the LVLM's integration of object identification and enumeration into mathematical problem-solving. Specifically, despite applying the CoT approach to demonstrate the two steps minimally required for solving VLMP (object identification and enumeration, followed by numerical operations), the LVLM's ability to integrate them for problem-solving remains limited, resulting in minimal performance improvement relative to vanilla prompting. This suggests a possible limitation in the

LVLM's ability to seamlessly combine these distinct skills despite its success in logical reasoning and accurate mathematical operations. Strikingly, we are able to show that this discrepancy occurs even when the LVLM correctly identifies and counts the objects when explicitly prompted in isolation to do so.

To overcome these limitations and close this multimodal reasoning gap, we propose a novel yet natural vision-and-language decomposed prompting method (henceforth: V&L DECOMP). This approach deconstructs the VLMP task into two subtasks: (i) a *vision-focused query task*: concentrates on identifying and enumerating objects within the image — specifically those that are relevant to the math problem — effectively grounding the problem in the visual context; and (ii) a *language-focused reasoning task*: leveraging the identified objects, this subtask employs textual reasoning to solve the mathematical problem, establishing a clear connection between the visual information and the solution process. Notably, this decomposition is made and streamlined explicitly, rather than by demonstration as in CoT. Testing on the Clevr-Math benchmark (Lindström and Abraham, 2022) and a VLMP-filtered subset of the VQA2.0 dataset (Goyal et al., 2017), shows that V&L DECOMP outperforms CoT (+7% Clevr-Math, +28% VQA2.0).

Looking ahead, we envision a "decompose by demonstration" approach for LVLMs inspired by our decomposition approach, subsequently formulating strategies and implementing an appropriate plan for solving a multimodal problem. Crucially, such a plan should allow the model to delegate specific tasks to specialized modalities in areas it lacks proficiency. Such strategies could potentially enhance multimodal problem-solving even further.

2 LVLMs and Prompting: Preliminaries

Large Vision and Language Models (LVMLs) The prevalent structure for large vision-language models (LVLMs, Li et al., 2023) is founded on a two-part backbone: (i) a Vision-Language Model (VLM) – this component serves as a bridge between visual and textual data, establishing a shared embedding space. Within this space, both images and their textual descriptions are represented as comparable points. Contrastive Language-Image Pre-training (CLIP, Radford et al., 2021) exemplifies this approach. (ii) a Large Language Model (LLM) – this component leverages both the language prompt and the image embedding generated by the VLM (part (i)). Despite the demonstrated capabilities of VLMs in object recognition and enumeration guided by a text-label (Jiang et al., 2023) and the remarkable capability of LLMs in solving MWP (Chen et al., 2022; Gaur and Saunshi, 2023; Yang et al., 2023; Imani et al., 2023), a performance gap remains in the VLMP task. As evidenced by MathVista (Lu et al., 2023), humans still outperform LVLMs by a margin of 15.5% on this task.

Prompting Strategies Chain-of-Thought (CoT) prompting is a recently developed technique that aims to improve the reasoning abilities of large language models (LLMs, Brown et al., 2020b; Chowdhery et al., 2023b) by explicitly encouraging them to explain their thought process step-by-step (Wei et al., 2022; Kojima et al., 2022). Program-of-Thoughts (PoT) builds upon the idea of CoT (Chen et al., 2022), using formal language elements like conditional statements and logical operators to guide the LLM through the problem-solving process. This allows the answer to delegate the computation steps to an external language interpreter, potentially simplifying implementation.

Both of these methods take a step-by-step approach, breaking the overall task into subtasks, and they both utilize demonstrations, where the model implicitly acquires knowledge of the task structure from observed solutions and explanations. However, the decomposition is done implicitly, and without explicitly planning the distinct subtasks at inference time. This enables knowledge transfer through demonstration sequences but leaves the concrete problem-solution structure implicit. In contrast, DECOMP (Khot et al., 2022) proposed for LLMs decomposes tasks by creating a planning prompt and then delegating the subtasks to different LLM handlers. Adapting these learn-by-demonstration-and-explanation approaches to multimodal tasks presents substantial challenges. Recent efforts demonstrated CoT application to only one benchmark, scienceQA (Zhang et al., 2023).

Contrary to such implicit learning from demonstrations and explanations, here we propose to explicitly decompose the prompting process, delegating the distinct tasks to the relevant task handlers. This entails segmenting a specific problem instance within the overall task into modality-specific subtasks. Subsequently, the input prompt is explicitly divided into smaller, targeted sub-prompts, each delegated to specialized handlers tailored to their

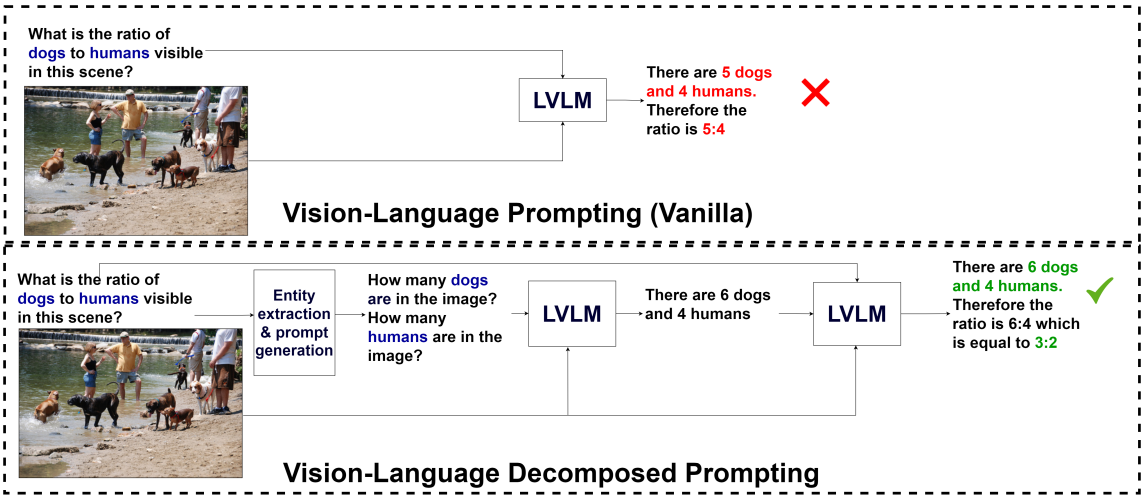


Figure 1: Illustration of a vanilla vision-language prompting compared to vision-language decomposed prompting.

respective modality. Furthermore, our approach builds on successive prompting (Dua et al., 2022), where the output of one subtask is passed to be used in prompting the handler in the next subtask.

3 The V&L DECOMP Approach

Due to the inherently multimodal nature of the VLMP task, requiring reasoning over both visual and textual information, we propose a vision-language decomposed prompting approach (V&L DECOMP, Figure 1). This approach breaks down the problem into two key sub-tasks, respective of the different modalities: (i) a *vision-focused query task* – focuses on identifying and enumerating objects relevant to the mathematical question within the image. This effectively grounds the problem in the visual context and extracts key elements for further processing. (ii) a *language-focused reasoning task* – leveraging the identified objects from the first sub-task, this stage employs textual reasoning to solve the mathematical problem. It utilizes the extracted visual information alongside the textual prompt to reach a solution through symbolic manipulation or mathematical reasoning processes.

In stage (i), we extract entities using an LLM (GPT-4) from the input question X = $x_0, x_1, ..., x_n$ where $\{x_i\}_{i=0}^{n}$ is a sequence of tokens, and the output is a set of entity spans from X, denoted as $E = \{e \mid e = (y_0, y_1, \ldots, y_m), 0 \leq m \leq n, \forall i \in \{0, 1, \ldots, m\}, \exists j, 0 \leq j \leq n, y_i = x_j\}$. We then generate a prompt querying about all the entities found in the question. P^0 = Concat({"How many e_i are in the image and what are they?")| $E_i \subseteq E, 0 \leq i \leq m$}). For example, E ={'blue objects','blocks'}, then P^0 = "How many blocks are in the image and what are they? How many blue objects are in the image and what are they?". In stage (ii), we pass the prompt P^0 and the image I to an LVLM and receive the sequence output Y^0 = LVLM(P^0, I). In the second stage, we concatenate the question X and the output of the first stage: $P^1 = Concat(Y^0, x)$. We then pass the image I and the prompt P^1 to an LVLM to receive the final answer Y^1 = LVLM(P^1, I).

4 Experiments

Goal We set out to empirically quantify the performance gap and subsequently evaluate the effectiveness of our proposed V&L DECOMP in addressing the performance gap on VLMP tasks.

Datasets To analyze the challenges in the VLMP task, we evaluate our method on two datasets: Clevr-Math (Lindström and Abraham, 2022), and arithmetic problems extracted from VQA2.0 (Goyal et al., 2017). We use a subset of 400 VLMPs from Clevr-Math (built on Clevr, Johnson et al., 2017). This selection covers 100 problems each in addition (Add.), subtraction (Sub.), adversarial (Adver.), and multi-hop subtraction. MathVista (Lu et al., 2023) extracted 188 math problems from VQA2.0. Out of this, we extracted 25 VLMPs that: (i) include arithmetic (addition, subtraction, multiplication, or division) (ii) are not strict counting problems or shape size comparisons; (iii) do not require world knowledge (e.g., "What is the MPG of an average city bus?"); and crucially, (iv) require *both* image and text to answer the problem.

	Test						Development				
	VQA2.0	Clevr-Math					Clevr-Math				
	All	All	Add.	Sub.	Adver.	Mult.	All	Add.	Sub.	Adver.	Mult.
Vanilla	0.28	0.51	0.63	0.49	0.48	0.44	0.52	0.75	0.51	0.43	0.39
CoT	0.32	0.50	0.64	0.53	0.51	0.33	0.50	0.69	0.55	0.37	0.38
V&L DECOMP	0.60	0.57	0.73	0.51	0.53	0.51	0.54	0.78	0.55	0.38	0.44

Table 1: Test and development sets accuracies for Clevr-Math and VQA2.0

	Clevr-Math										VQA2.0									
	CA		correct isolated OR and WA		correct math in CA		wrong OR in CA		wrong OR in WA		CA		correct isolated OR and WA		correct math in CA		wrong OR in CA		wrong OR in WA	
	c	μ	c	μ	c	μ	c	μ	c	μ	c	μ	c	μ	c	μ	c	μ	c	μ
Vanilla	37	0.46	14	0.39	37	1	11	0.30	42	1	7	0.28	8	0.44	7	1	0	0	18	1
CoT	35	0.44	11	0.31	35	1	5	0.14	44	0.98	8	0.32	8	0.47	8	1	0	0	17	1
V&L DECOMP	39	0.49	0	0.00	39	1	3	0.08	41	1	15	0.60	1	0.10	15	1	1	0.07	10	1

Table 2: Analyzed predictions on 80/25 Clevr-Math/VQA2.0 samples, showing accuracy (μ) and count (c) for correct (CA) and incorrect responses (WA); correct mathematics, and incorrect object recognition (OR) in responses.

Models We leveraged the OpenAI's multimodal GPT-4 model with vision capabilities (GPT-4V)[1] and GTP-4 model version 0613, with temperature 0 in both models for consistent text generation.

5 Results

Table 1 presents the performance of three prompting strategies on the VQA2.0 and Clevr-Math datasets. Notably, our V&L DECOMP approach achieves accuracy gains of 7% and 28% over CoT on Clevr-Math and VQA2.0, respectively. This disparity in improvement can be attributed to the differing mathematical complexities of the datasets. VQA2.0 poses a greater challenge due to its inclusion of multiplication, division, and larger numbers (tens of thousands), requiring intricate mathematical reasoning from the model.

Table 2 reveals that even when models correctly identify and count objects in an isolated prompt, they still provide incorrect answers (WA) to the main VLM task when prompted again on the same image, mainly in CoT and Vanilla prompting strategies. Notably among all approaches, our V&L DECOMP approach performs best in mitigating this phenomenon. Table 2 also reveals that inaccurate object recognition was a major factor contributing to incorrect responses across all strategies. This was further confirmed by an oracle experiment conducted on the Clevr-Math validation set. In this experiment, GPT-4V was provided with the ground-truth object labels from the image annotations. This intervention resulted in a significant improvement in model accuracy, reaching 79.25%. Additionally, more objects in the images significantly hinder ($p < 0.001$) LVLM's performance in the Clevr-Math validation set, as shown by a strong negative Spearman's ranking correlation (-0.92, Spearman, 1961). These findings highlight the critical role of robust object recognition in achieving high performance on VLMP tasks. Furthermore, these findings, coupled with the superior performance of our V&L DECOMP prompting strategy, suggest that decomposing the problem by addressing object recognition as a separate subtask, potentially through dedicated VLMs trained specifically for this purpose, could be a promising avenue for future research.

6 Conclusion

This study has demonstrated the potential of the novel vision-language decomposed prompting method (V&L DECOMP) in addressing the challenges of Vision and Language Math Problems (VLMP). By deconstructing the VLMP task into two focused subtasks, respective of the different modalities, we have been able to significantly improve the performance of large vision-language models (LVLMs) in streamlining the object identification and enumeration into mathematical problem-solving. The V&L DECOMP approach has shown a marked increase in accuracy on two datasets, outperforming the chain-of-thought (CoT) approach. These findings underscore the importance of modal-specific decompositions in enhancing the reasoning capabilities of LVLMs, paving the way for solving more complex multi-modal reasoning tasks in the future, and leaving ample space for researching such decomposition learning by demonstration.

[1] https://platform.openai.com/docs/guides/vision

Limitations

Limited VLMP Datasets The current scarcity of visual and language math problem (VLMP) datasets has constrained the scope of experimental analysis in this research. While MathVista (Lu et al., 2023) categorizes three VLMP datasets: IconQA (Lu et al., 2021), TabWMP (Lu et al., 2022b), and Clevr-Math (Lindström and Abraham, 2022); only Clevr-Math contains VLMPs that require arithmetic operations and present problems over natural images. IconQA contains less than a handful of VLMP problems that go beyond counting and comparisons, whereas TabWMP contains questions over images of tables only, with no natural images. Although Clevr-Math contains VLMPs, approximately 5% of the data contains incorrect labels. Therefore, we implemented a manual curation process to rectify these inconsistencies within a representative subset (400 samples) of the test set. This approach, while mitigating the impact of label noise, is inherently costly and time-consuming and limits the scale of results we can provide, compared with the full dataset. Alternative strategies, such as synthetic question-answer pair generation or crowdsourcing data, are beyond the scope of the present (short focused) contribution and we leave it for future research.

Decomposition with Expert Models The correlation shown between incorrect answers and inaccurate object recognition in Section 5, suggests that object recognition remains a significant hurdle in VLMP tasks. V&L DECOMP possesses a unique advantage in addressing this, as it partitions the task into subtasks. Theoretically, the first stage could utilize a specialized VLM adept at object counting. However, at present, we have not identified a VLM that surpasses GPT-4V. We conducted a quantitative analysis on 30 randomly selected images from the COCO dataset (Lin et al., 2014), which contain between 2-10 objects. When comparing the performance of Blip2 (Li et al., 2023) and GPT-4V, we found that GPT-4V was superior. Exploring the potential of integrating models that specialize in different aspects of the task, is left for future work.

Explicit vs. Implicit Decomposition While our V&L DECOMP approach demonstrates improved performance in the VLMP task, its scalability raises concerns given the pre-defined task planning (although instance-specifics details are handled automatically at inference time). This contrasts with known methods for learning-by-demonstration-and-explanation, which are significantly easier to create and are readily producible — and yet, as we empirically show here, fall surprisingly short in their efficacy to handle the complications of multimodal reasoning. Therefore, future research should explore effective methods for learning-by-demonstrations beyond simple CoT approaches, for modality-aware decompositions in the vision-and-language domains, while incorporating lessons learned from the V&L DECOMP approach.

References

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov,

and Noah Fiedel. 2023a. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023b. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. 2022. Successive prompting for decomposing complex questions. *arXiv preprint arXiv:2212.04092*.

Vedant Gaur and Nikunj Saunshi. 2023. Reasoning in large language models through symbolic math word problems. *arXiv preprint arXiv:2308.01906*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. Mathprompter: Mathematical reasoning using large language models. *arXiv preprint arXiv:2303.05398*.

Ruixiang Jiang, Lingbo Liu, and Changwen Chen. 2023. Clip-count: Towards text-guided zero-shot object counting. *arXiv preprint arXiv:2305.07304*.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. 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.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Nate Kushman, Yoav Artzi, Luke Zettlemoyer, and Regina Barzilay. 2014. Learning to automatically solve algebra word problems. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 271–281.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Adam Dahlgren Lindström and Savitha Sam Abraham. 2022. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating math reasoning in visual contexts with gpt-4v, bard, and other large multimodal models. *arXiv e-prints*, pages arXiv–2310.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pages 2507–2521. Curran Associates, Inc.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022b. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. *arXiv preprint arXiv:2209.14610*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.

Charles Spearman. 1961. The proof and measurement of association between two things.

Taylor Webb, Keith J Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. 2023. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

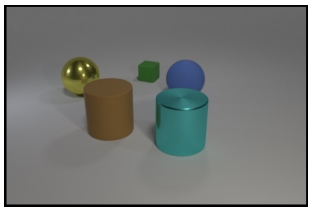
A CoT Prompting Examples

From VQA2.0 *You are asked to answer a mathematical <question> about an image. Before answering the question, here are examples of question-answer pairs. Use this approach when answering the <question>. Example 1: If we remove one lamp from the room in the image how many lamps will be left?*

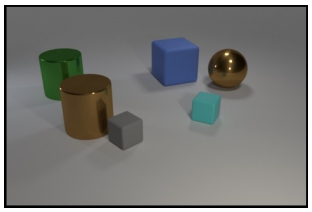
There are 2 lamps in the image. Any additional lights are merely mirror reflections. Then if we remove 1 lamp there will be only one lamp. Mathematically, this is represented as 2-1=1
Example 2: What is the ratio of women to men?

There are 3 women and 2 men in the image. To find the ratio, we divide the number of women by the number of men, resulting in 3/2.

From Clevr-Math *You are asked to answer a mathematical <question> about an image. Before answering the question, here are examples of question-answer pairs. Use this approach when answering the <question>. Example 1: Subtract all brown things. Subtract all brown cylinders. How many objects are left?*
a. There are 2 brown things in the image: 1 big brown metal cylinder and 1 big brown metal sphere.



b.There is 1 brown cylinder in the image: 1 big brown metal cylinder. c. There are 6 objects in the image in total: 1 big brown metal cylinder, 1 big green metal cylinder, 1 gray small rubber cube, 1 cyan small rubber cube, 1 big blue rubber cube, 1 big brown metal sphere. d. After subtracting the 2 brown things we are left with 4 objects in the image. We don't need to subtract the brown cylinder again as we already subtracted it as part of 'brown things'. My answer is: 4. Another example: Add two cyan blocks. How many cyan objects are there?



a. There is 1 cyan object in the image: 1 large cyan metal cylinder. b. After adding 2 cyan blocks, there are 3 cyan objects in the image. My answer is: 3.

B Additional Results

The Object Recognition Challenge All three prompting strategies exhibit a strong, negative correlation between success rate and the number of objects in the image, as shown in Table 3. Each strategy's Spearman's rank correlation coefficient is less than -0.85 and statistically significant ($p < 0.05$), indicating that increased object count consistently hinders model performance across all strategies.

	Correlation	p-value
Vanilla	-0.92	0.001
CoT	-0.85	0.007
V&L DECOMP	-0.91	0.002

Table 3: Spearman's ranking correlation between success rate and the number of objects in the image.

Error Analysis Building upon the initial analysis, Table 4 dives deeper into error patterns across

	Incorrect VLMP answers						correct VLMP answers with errors							
	Vanilla		CoT		V&L DECOMP		Vanilla		CoT		V&L DECOMP			
	c	μ	c	μ	c	μ	c	μ	c	μ	c	μ	**Example**	**Explanation**
Total	43	0.54	45	0.56	41	0.51	1	0.3	5	0.14	3	0.08	NA	NA
Counting	30	0.70	27	0.60	29	0.70	7	0.64	4	0.8	3	1	Answer: There are 10 objects in total.. Question: ...How many blue blocks are left? Answer: ...there is only 1 small blue metal/shiny cube...	Counted 10 objects instead of 9
Hallucination	19	0.44	20	0.44	13	0.31	8	0.73	3	0.6	1	0.33	Question: Subtract all small cyan matte cylinders. Subtract all cyan cylinders. Answer: Subtracting the small cyan matte cylinder and the cyan cylinder, there is 1 object left.	There are no blue cubes in the image.
Reasoning	1	0.02	2	0.04	1	0.02	1	0.33	0	0	0	0		Subtracted the same object twice

Table 4: Error Analysis for Different Prompting Strategies in Clevr-Math Validation Set.

the three prompting strategies. Using an 80-sample study from the Clevr-Math validation set, the analysis categorizes errors into three key phenomena: (i) *Counting errors* – Inaccurate object quantification within the image. (ii) *Hallucinations* – Introduction of non-existent objects into the model's interpretation. (ii) *Reasoning errors* – Faulty logical processes applied by the model, such as double-subtracting the same object. Analysis of error types reveals that V&L DECOMP exhibits the lowest frequency of explanatory errors, regardless of whether the final answer is correct or incorrect. Counting errors were the most prevalent, followed by 'hallucinations'.'Reasoning' errors are relatively negligible. Notably, V&L DECOMP consistently demonstrated the least susceptibility to hallucinations across both answer categories. Conversely, vanilla prompting yielded the highest frequency of counting errors, regardless of answer correctness.