SELF-IMAGINE: EFFECTIVE UNIMODAL REASONING WITH MULTIMODAL MODELS USING SELF-IMAGINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The potential of Vision-Language Models (VLMs) often remains underutilized in handling complex text-based problems, particularly when these problems could benefit from visual representation. Resonating with humans' ability to solve complex text-based problems by (1) creating a visual diagram from the problem and (2) deducing what steps they need to take to solve it, we propose SELF-IMAGINE. We leverage a single Vision-Language Model (VLM) to generate a structured representation of the question using HTML, then render the HTML as an image, and finally use the same VLM to answer the question using both the question and the image. Our approach does not require any additional training data or training. We evaluate our approach in three mathematics tasks and nine general-purpose reasoning tasks using state-of-the-art (LLAVA-1.5 and GEMINI PRO) VLMs. Our approach boosts the performance of VLM on all math tasks (on average GSM8K: +3.145%; ASDIV: +3.25%; SVAMP: +6.90%) and the majority of the general-purpose reasoning tasks by 3.20% to 6.00% on average.

1 INTRODUCTION

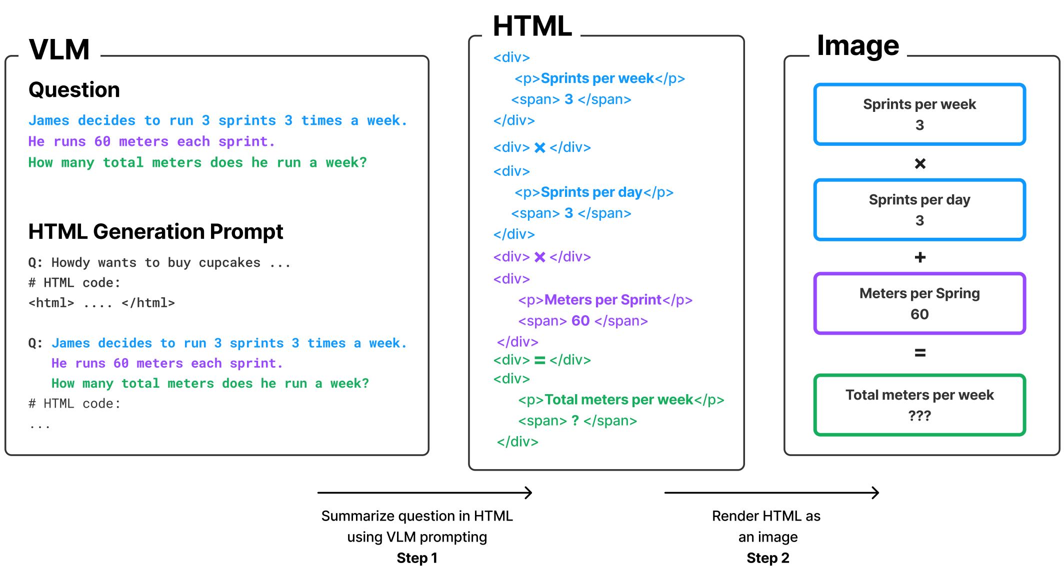


Figure 1: Generating an image from a question via a single VLM through HTML.

Vision Language Models (VLM) are getting increasingly adept at solving a wide range of reasoning tasks (Liu et al., 2023a;b; You et al., 2023; Ye et al., 2023; Chen et al., 2023b; Zhang et al., 2023; Chen et al., 2023a; Dai et al., 2023; Lu et al., 2023). As these capabilities advance, VLMs are set to replace the current text-only language models for general-purpose interfaces like BARD (GoogleAI, 2023) and ChatGPT (OpenAI, 2021). In such scenarios, the deployed VLM would be required to handle a wide variety of end-user queries. Crucially, this includes queries that are not inherently multimodal, such as math-reasoning problems or program synthesis (Cobbe et al., 2021).

A key question arises in these situations: How should a VLM, capable of functioning in a text-only mode like a Language Language Model (LLM), handle text-based queries? While the default approach is to process these queries purely as text, this method does not fully exploit the VLM's capabilities in image processing. Recent studies on human problem-solving provide a clue to addressing this gap: humans often draw visual representations to better understand and solve problems (Boonen et al., 2014; van Garderen et al., 2012; Krawec, 2014).

Building on this insight, we propose SELF-IMAGINE–a technique designed to enhance the reasoning abilities of VLMS on text-only tasks through visualization (Figure 1). SELF-IMAGINE initially generates a graphical representation of the text query using the VLM. Then, the *same* VLM is used to solve the problem using both the original question and the self-generated image.

An inherent challenge is that advanced VLMs are not typically equipped for direct image generation. To circumvent this, we utilize the VLM's code generation capabilities to generate HTML code visually representing the query information. This HTML is then rendered as an image, which, when used in conjunction with the original text query, allows the VLM to operate with both textual and visual information. With SELF-IMAGINE, the VLM efficiently serves dual purposes: generating visual representations and solving the problem. This strategy effectively reduces reliance on separate image generation models such as DALL-E (Shi et al., 2020), streamlining the problem-solving process.

We test our approach across three mathematical reasoning tasks and nine general-purpose reasoning tasks. We find that SELF-IMAGINE is particularly effective when the generated image demonstrates the information in a structured way that corresponds to the reasoning steps needed to be performed to solve the question. We show that SELF-IMAGINE improves the performance of state-of-the-art VLMs (LLAVA-1.5 (Liu et al., 2023a), GEMINI PRO (Team, 2023)) across all math reasoning tasks namely GSM8K (Cobbe et al., 2021) (+1.67 to 4.62%), ASDIV (Miao et al., 2020) (+2.01 to 4.49%) and SVAMP (Patel et al., 2021) (+4.5 to 9.30%), and achieves superior performance (ranging from 0.40% to 13.2% improvement) in five out of nine general-purpose reasoning tasks while receiving comparable accuracy to question only setup in other tasks.

2 SELF-IMAGINE

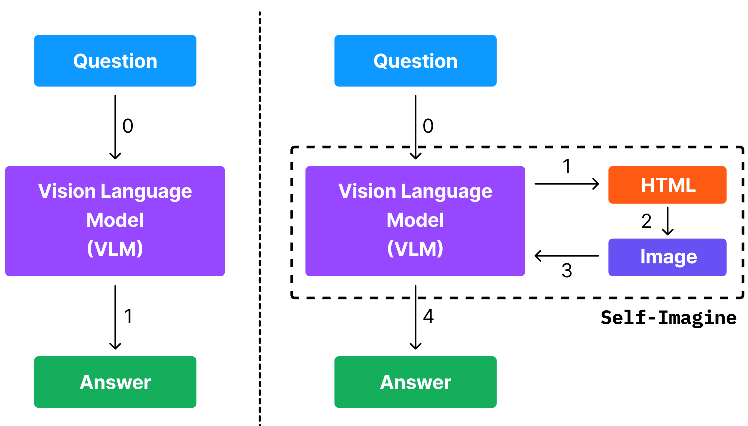


Figure 2: [**Left**] **Reasoning using VLM without SELF-IMAGINE:** Given a question (0), the VLM generates an answer (1). [**Right**] **Reasoning using VLM with SELF-IMAGINE**: Given a question (0), the VLM generates a structured representation of the question using HTML (1). The HTML is rendered as an image (2) which is then passed along with the question to the VLM again (3). The VLM finally generates the answer by combining both vision and language modalities (4).

Unlike Large Language Models (LLM), Vision Language Models (VLM) can combine multiple modalities in the same representation space and perform complex reasoning. However, when it comes to unimodal downstream tasks (e.g., math-reasoning), VLMs are not fully leveraged due to the absence of additional modalities. In SELF-IMAGINE, we circumvent this by generating a visual representation for a given reasoning question using the VLM in the form of an image. Then, the same VLM is fed both the question and the generated image to answer the question. In the following section, we expand on the image generation from the question.

2.1 GENERATE IMAGE FROM QUESTION

While VLM cannot generate images directly, they are pre-trained on large corpus of programs and thus are proficient in code generation. Thus, we utilize code generation capabilities of these models to create an image for the question. While there are several choices for choosing a representation (SVG (St.Laurent et al., 2001), Tikz (Tantau, 2022)), we use HTML due to its prevalence and its ability to easily generate structured information from questions using tables, lists, flow charts, etc.

Generate HTML from Question. To convert natural language questions into HTML, we choose two Vision Language Models (VLM): LLAVA-1.5 (Liu et al., 2023a) & GEMINI PRO (Team, 2023)), due to their impressive performance on a wide range of reasoning tasks. Since multimodal models are not traditionally trained for HTML generation, we approach this using a few-shot prompt, interleaving natural language questions with HTML codes. For each natural language question q_i, we generate a corresponding HTML code h_i. These are paired as $\langle q_i, h_i \rangle$ to form a prompt $p = \{q_j, h_j\}_{j=1}^{K}$, where $K = 5$ represents the number of in-context examples chosen for diversity in reasoning tasks. Given a new question q_t, we combine it with the prompt p and a placeholder image I_d, and input these into the VLM to generate the HTML h_t for q_t as shown in Equation 1.

$$h_t = \text{VLM}(p \,||\, q_t, I_d) \tag{1}$$

Convert HTML to Image. After generating HTML from questions, we use the 'imgkit' python library to render these HTML codes into images. To evaluate the role of images in reasoning tasks, we conduct experiments both with and without the generated images. We append task-specific prompts to the questions, as detailed in Table 3. In the image-inclusive experiments, we use the HTML-generated images alongside the concatenated prompts and questions, inputting these into the VLM for processing.

$$\begin{aligned} I_g &= f(h_t) \\ y_t &= \text{VLM}(p||q_t, I_g) \end{aligned} \tag{2}$$

Here, f represents the HTML renderer, and I_g represents the final generated image from the question. y_t is the answer generated using the question with the prompt $(p||q_t)$ and the image (I_g).

3 EXPERIMENTAL SETUP

Benchmarks. We explore two kinds of reasoning tasks to evaluate our approach: (1) *math word problems* consisting of GSM8K (Cobbe et al., 2021), ASDIV (Miao et al., 2020), and SVAMP (Patel et al., 2021) and (2) symbolic reasoning consisting of NAVIGATE, GEOMETRIC SHAPES, TRACKING SHUFFLED OBJECTS, PENGUINS IN A TABLE, COLORED OBJECTS, DATE UNDERSTANDING, and OBJECT COUNTING tasks from BIG-Bench Hard (Suzgun et al., 2022).

Baselines. For the baseline, we consider zero-shot prompting where we only pass a basic prompt (Table 3) and the question. We performed greedy decoding from the language model using a temperature of 0. Note that this is a realistic setup for current open-source multimodal LLMs, which cannot accept a prompt interleaved with text and images.

Vision Language Models. We use LLaVA-1.5 (Liu et al., 2023a) and GEMINI PRO (Team, 2023) as our VLMs and keep each one of them consistent throughout the HTML generation phase to the question-answering phase. LLaVA-1.5 uses a CLIP ViT-L (Radford et al., 2021) as a vision encoder and Vicuna 13B (Chiang et al., 2023) as the LLM. Conversely, GEMINI PRO is built on Transformer architecture (Vaswani et al., 2017) and is trained with a wide range of multimodal data. The architecture of this model has not been disclosed yet. In this paper, we accessed GEMINI PRO through Google AI Studio. GEMINI PRO comes with default safety features that block certain questions, especially those involving potentially illegal or sensitive content. For our analysis, we disabled these safety settings.

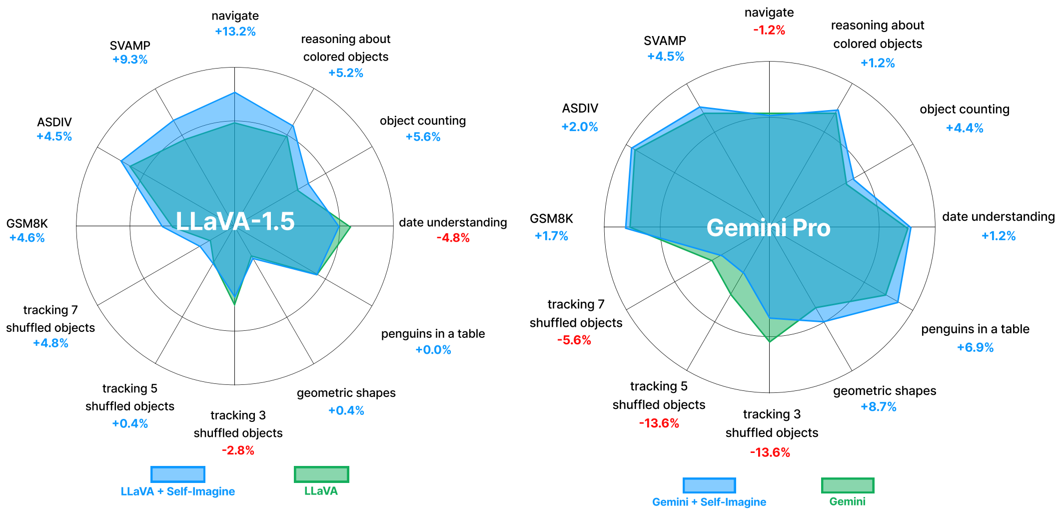


Figure 3: **SELF-IMAGINE main results**: SELF-IMAGINE improves accuracy over a diverse range of mathematical and symbolic reasoning tasks.

Evaluation During the evaluation, we slightly modified the standard evaluation protocol (Gao et al., 2023a), which consisted of matching the words "The answer is" followed by a numerical output. We found that the VLM sometimes fails to follow this sentence verbatim even when it produces the correct answer. To accommodate these cases, we simply take the last number/option of the generated text as the answer to the question.

4 RESULTS

We summarize our results across three math and nine reasoning tasks in Table 1. We define the baseline setup as *'Question Only'* when we only feed the question with the basic prompt to the VLM. SELF-IMAGINE is indicated by the *'Question + Image'* setup where we generate the HTML from the question at first and pass the rendered image from HTML along with the basic prompt and question to the VLM as input (Equation 2).

SELF-IMAGINE improves the VLMs' performance in all math reasoning tasks: for example, SELF-IMAGINE improves the base LLAVA-1.5 and GEMINI PRO by 9.30% and 4.50% accordingly in SVAMP. In OBJECT COUNTING (LLAVA-1.5: +5.60%; GEMINI PRO: +4.40%), COLORED OBJECTS (LLAVA-1.5: +5.20%; GEMINI PRO: +1.20%) and GEOMETRIC SHAPES (LLAVA-1.5: +0.40%; GEMINI PRO: +8.70%), inclusion of SELF-IMAGINE improves both VLMs.

LLAVA-1.5 and GEMINI PRO have different subsets of symbolic reasoning tasks in which they benefit from SELF-IMAGINE. In particular, LLAVA-1.5 benefits from SELF-IMAGINE in tasks involving multiple variables e.g., navigation and tracking multiple objects tasks, as the image provides additional structured information on top of the question. On the contrary, GEMINI PRO + SELF-IMAGINE excels in list and tabular reasoning tasks such as DATE UNDERSTANDING (+1.20%) and PENGUINS IN A TABLE (+6.85%). All these tasks require diverse reasoning abilities, and the improvement across these tasks represents the generalizability of SELF-IMAGINE.

However, SELF-IMAGINE hurts the performance of VLMs in some of the symbolic reasoning tasks - for LLAVA-1.5: DATE UNDERSTANDING (-4.80%) and TRACKING SHUFFLED OBJECTS of three objects (-2.80%); for GEMINI PRO: NAVIGATE (-1.20%), TRACKING SHUFFLED OBJECTS of three objects (-13.6%), of five objects (-13.6%), and seven objects (-5.60%). These tasks are easier to solve using only the question rather than having an image. The reason behind degradation with an image is two-fold: (1) the generated images are incorrect and visually not informative given the question (DATE UNDERSTANDING, NAVIGATE), (2) HTML cannot visually portray terms like swap between objects and cannot keep track of an object after multiple swaps (TRACKING

SHUFFLED OBJECTS). These results indicate that stronger image generation capabilities that capture consecutive progression of reasoning might help to boost the performance of the VLM.

In the following section, we demonstrate that the improvement is highly correlated with the quality of the generated image, underscoring the dependency on the ease of converting text into an image. In addition, an image that appropriately captures the flow of reasoning always guides the VLM to the correct reasoning path.

Task	**Dataset**	LLAVA-1.5		GEMINI PRO	
		Question Only	Question + Image	Question Only	Question + Image
Math Reasoning	GSM8K	26.69	**31.31** (+4.62)	74.37	**76.04** (+1.67)
	ASDIV	52.24	**56.73** (+4.49)	82.01	**84.02** (+2.01)
	SVAMP	43.50	**52.80** (+9.30)	69.50	**74.00** (+4.50)
Symbolic Reasoning	OBJECT COUNTING	31.20	**36.80** (+5.60)	46.40	**50.80** (+4.40)
	NAVIGATE	44.80	**58.00** (+13.2)	**60.80**	59.60 (-1.20)
	COLORED OBJECTS	44.80	**50.00** (+5.20)	70.40	**71.60** (+1.20)
	DATE UNDERSTANDING	**50.00**	45.20 (-4.80)	72.80	**74.00** (+1.20)
	PENGUINS IN A TABLE	41.10	41.10 (0.00)	70.55	**77.40** (+6.85)
	GEOMETRIC SHAPES	**14.40**	14.80 (+0.40)	48.00	**56.70** (+8.70)
	TRACKING SHUFFLED OBJECTS (3)	**33.60**	30.80 (-2.80)	**60.00**	46.40 (-13.6)
	TRACKING SHUFFLED OBJECTS (5)	18.00	**18.40** (+0.40)	**41.20**	27.60 (-13.6)
	TRACKING SHUFFLED OBJECTS (7)	12.00	**16.80** (+4.80)	**34.40**	28.80 (-5.60)

Table 1: Comparison of accuracy between 'Question Only' and 'Question + Image' across diverse reasoning tasks where the image has been generated using SELF-IMAGINE.

5 ANALYSIS

5.1 MATH REASONING

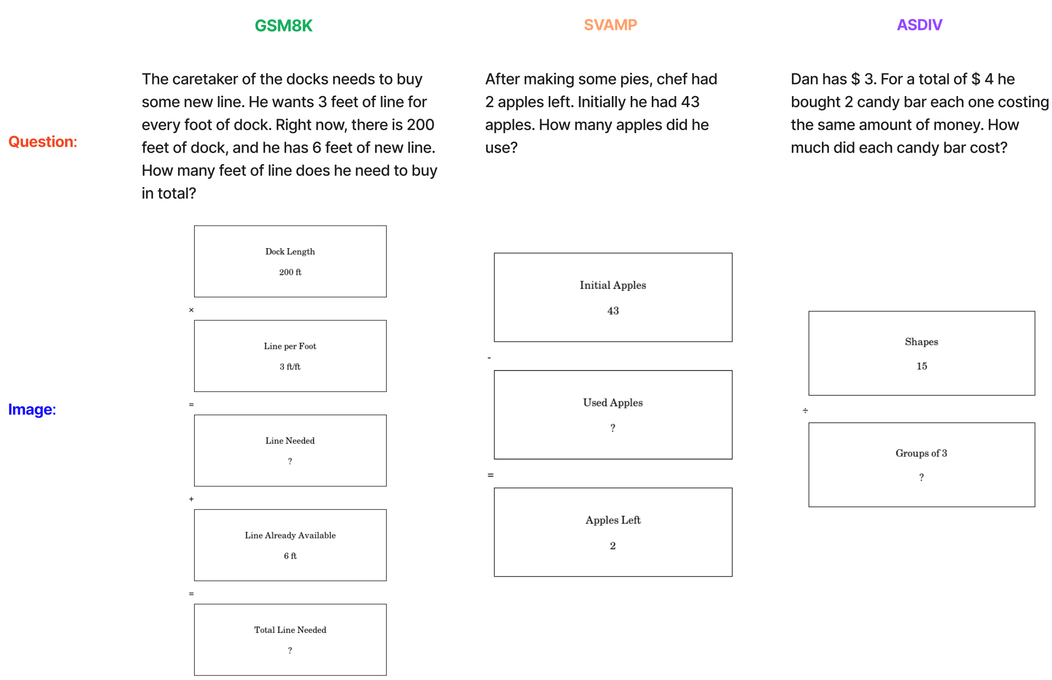


Figure 4: Example from math world problem tasks.

For math reasoning tasks, we analyze the performance of VLMs with and without image support. This analysis includes examining performance variations across question complexity, the length of the reasoning chain, and specific instances where images contribute positively or negatively to problem-solving. The generated images, as depicted in Figure 4, predominantly feature boxes, each labeling a variable and its value, designed to simplify and clarify the information presented in the question.

Why does image help? The primary advantage of using images lies in their ability to distill complex information into a more manageable format. In several tasks, particularly those involving substantial irrelevant data (e.g., GSM8K, ASDIV), an image serves as a focused reference point, enabling the model to concentrate on key variables and their values (see Table 2, Table 4 for examples). Additionally, images often include variable names marked with question marks, as shown in Figure 4, which guide the model in identifying the critical elements necessary for multi-step reasoning.

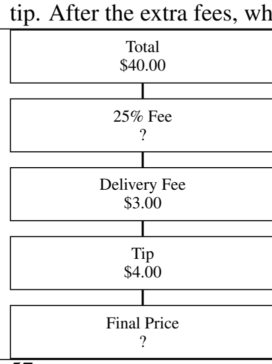
Question	Stephen placed an online order for groceries. His final bill came to $40.00. Because this was through a delivery vendor, they tacked on a 25% fee to his final total and charged him $3.00 in delivery fees. Stephen also added a $4.00 tip. After the extra fees, what was the final price of Stephen's groceries?
Image	
Label	57
Output With Image	To find the final price of Stephen's groceries, we need to calculate the total cost after the extra fees and tip have been added. 1. Calculate the 25% fee: $40.00 (final bill) * 0.25 = $10.00 (25% fee) 2. Add the delivery fee to the final bill: $40.00 (final bill) + $10.00 (25% fee) + $3.00 (delivery fee) = $53.00 (total cost with fees) 3. Add the tip to the total cost: $53.00 (total cost with fees) + $4.00 (tip) = $57.00 (final price of groceries) The answer is **$57.00**.
Output Without Image	The final price of Stephen's groceries was $40.00 + 25% fee + $3.00 delivery fee + $4.00 tip = $40.00 + 0.25*40 + 3 + 4 = $40.00 + 10 + 3 + 4 = $40.00 + 13 = $53.00. The answer is **$53.00**.

Table 2: Example of Image improving reasoning in GSM8K task for LLAVA-1.5.

Image helps solve moderately complex questions. In general, longer questions tend to be complex. Here, we examine the performance variation regarding question length as detailed in Figure 6. We find that image helps LLAVA-1.5 more than GEMINI PRO in longer and more complex questions in ASDIV and SVAMP tasks. This finding aligns with the previous explanation, i.e., the image removes unnecessary verbose from the question, making the reasoning process easier.

However, we can also observe that for more complex questions in the GSM8K task (question length > 70 for LLAVA-1.5 & question length > 50 for GEMINI PRO), performance with images deteriorates compared to performance without images. This decline stems from the inadequate HTML generated by longer questions, which often fail to encapsulate all the necessary information. Therefore, images generated from those HTMLs confuse the VLMs rather than help.

This observation also holds for questions with longer reasoning chains depicted in Figure 7 for the GSM8K task. Questions that require a longer chain-of-thoughts (COT) are not better represented with images for LLAVA-1.5. However, GEMINI PRO is robust to increasing COT length and rather benefits from having a structured representation for complex questions. This analysis also presents an opportunity for future research. It suggests that the most challenging questions, which intuitively could benefit the most from the structural and contextual support provided by images, are precisely where current methodologies for image generation fall short.

Why does the image hurt? While images generally enhance the VLM's reasoning, specific scenarios lead to diminished performance. A notable issue arises during HTML generation, where the VLM occasionally pre-solves arithmetic sequences, embedding them into the image (Table 5). This can mislead the model if the embedded calculations are incorrect. Furthermore, certain concepts like 'trade/exchange' or 'add/delete' are challenging to represent visually, leading to inaccuracies in questions involving these terms. Another complication involves questions with fractions, such as 'Shelly ate 3/4 of the eggs from a dozen.' The corresponding images often depict these fractions in a simplified form (e.g., a box labeled 'Already ate: 3/4 $\times$ 12'), which the model struggles to compute accurately as it requires the execution of multiple operations (i.e., division and multiplication) simultaneously. Similarly, when the VLM tries to execute multiple operations mentioned in the image, it usually generates the incorrect answer. For example, in Table 9, with the image, the VLM executes four operations in a single line (i.e., $10 * 1/2 + 15 * 1/3$) and ends up generating the wrong answer. But without an image, the calculation is broken down even further, producing the correct answer. This problem might be solved with an improved image that breaks down each step as a single operation consisting of two numbers.

5.2 SYMBOLIC REASONING

In this category, we focus on nine diverse reasoning tasks from BIG-Bench Hard benchmark (Suzgun et al., 2022) to observe the importance of image. We break down the overall accuracy by tasks and analyze the performance by question complexity and answer types. The images generated with HTML for the tasks are images with labeled/colored boxes (Figure 5b), tables (Figure 5a, Figure 5c). Occasionally, we find that the generated image simply contains the text (as in Table 10).

Why and when does image help? The overall accuracy indicates a decent improvement for LLAVA-1.5 (2.56%) with SELF-IMAGINE (as Figure 3) where GEMINI PRO receives slight accuracy loss (-1.69%) with self-generated image. We further break down the results across the tasks. As shown in Figure 3, adding an image augments the performance of LLAVA-1.5 in the majority of symbolic reasoning tasks while achieving comparable performance in others. In parallel, adding images improves GEMINI PRO in tasks that require shape, color, list, and tabular reasoning such as COLORED OBJECTS, OBJECT COUNTING, DATE UNDERSTANDING, PENGUINS IN A TABLE, and GEOMETRIC SHAPES.

For COLORED OBJECTS, PENGUINS IN A TABLE, and OBJECT COUNTING tasks, the VLMs generate well-structured tables or multiple boxes in rows with variable names in one column and corresponding values in another column. Thus, when solving with an image, the reasoning problem simplifies to finding column sums or specific table elements. Notably, GEMINI PRO, being a decent table parser (Akter et al., 2023), excels in these tasks with images. In GEOMETRIC SHAPES, the HTML simply depicts the shape provided in the SVG vector. As a result, image helps both VLMs by providing a visual reference of the intended shape in the question (as Table 8).

In contrast, TRACKING SHUFFLED OBJECTS requires tracking multiple objects through consecutive swaps. As mentioned in subsection 5.1, VLMs struggle to depict swaps between objects using HTML (Table 6), leading to degradation in performance with the image in TRACKING SHUFFLED OBJECTS tasks. However, TRACKING SHUFFLED OBJECTS of five and seven objects tasks are notably more complex than the three objects task as they require tracking more objects and their swaps and text-only LLAVA-1.5 performs poorly in these tasks. Having an image that logs each object attribute and their swaps, helps LLAVA-1.5 to solve the question accurately rather than having no image. Hence, we can see the improvement of LLAVA-1.5 with the increasing number of objects for TRACKING SHUFFLED OBJECTS tasks.

Finally, in NAVIGATE task, LLAVA-1.5 significantly improves with image inclusion, while GEMINI PRO shows little degradation in accuracy. Unlike other tasks, the NAVIGATE task is challenging to depict using HTML. Therefore, most of the images generated with both VLMs for this task contain texts either showing the question or necessary reasoning steps in natural language (Table 10). Without an image, LLAVA-1.5 performs poorly compared to GEMINI PRO on this task. However, with images, the LLAVA-1.5 executes additional reasoning during HTML generation, thereby increasing the likelihood of predicting the correct answer in the presence of an image. This phenomenon also explains GEMINI PRO's improvement in the DATE UNDERSTANDING task with images, as the generated HTML primarily offers reasoning steps in natural language.

Question: I have a table, and three cars. How many objects do I have?

Image:

Objects Count	
Table	1
Cars	3
Total Objects	?

(a) OBJECT COUNTING

Question: On the floor, you see a bunch of objects arranged in a row: a fuchsia jug, a black teddy bear, a blue envelope, a silver paperclip, a teal pencil, and a yellow keychain. What is the color of the object furthest from the silver object?

Image:

Fuchsia	Black	Blue	Silver	Teal	Yellow
Jug	Teddy bear	Envelope	Paper clip	Pencil	Keychain

(b) COLORED OBJECTS

Question: Here is a table where the first line is a header and each subsequent line is a penguin: name, age, height (cm), weight (kg) Louis, 7, 50, 11 Bernard, 5, 80, 13 Vincent, 9, 60, 11 Gwen, 8, 70, 15 For example: the age of Louis is 7, the weight of Gwen is 15 kg, the height of Bernard is 80 cm. We now add a penguin to the table: James, 12, 90, 12 How many penguins are less than 8 years old and weight more than 12 kg?

Image:

Name	Age	Height (cm)	Weight (kg)
Louis	7	50	11
Bernard	5	80	13
Vincent	9	60	11
Gwen	8	70	15
James	12	90	12

(c) PENGUINS IN A TABLE

Figure 5: Examples from some BIG-Bench Hard sub-tasks.

Image helps with shorter (GEMINI PRO) and more complex questions (LLAVA-1.5). Following subsection 5.1, we investigate the impact of the image in the reasoning process with increasing question length. Here, we observe distinct behaviors in two VLMs. As depicted in Figure 8, LLAVA-1.5 benefits from images with both simpler, shorter questions and more complex ones, while GEMINI PRO's performance declines as question length increases. Generating high-quality HTML is also easier for simpler and shorter questions, which benefits both VLMs during question answering with the appropriate image. However, with longer questions, the generated HTMLs tend to ignore some information or can not summarize all information in a structured manner. This results in lower performance compared to without image setup. Interestingly, Figure 8 shows higher performance for LLAVA-1.5 in with-image setup when the question length exceeds 120 words. This category includes the TRACKING SHUFFLED OBJECTS of seven objects task, which requires tracking seven objects during multiple swaps. As mentioned earlier, the image depicts the role of each object in the question and their swaps, which simplifies the reasoning process. Unlike LLAVA-1.5, GEMINI PRO particularly tries to solve the swaps step-by-step in the HTML rather than just logging the information. However, it fails to keep track of the objects after multiple swaps - resulting in the generation of an incorrect HTML and a dramatic drop in accuracy with images.

Why does the image hurt? Despite the benefits observed in certain tasks, incorporating images into the reasoning process can worsen performance in others. We observe that the reason behind the performance drop-off is two-fold: (1) images generated from HTML are incorrect or missing information, and (2) generated images cannot depict the reasoning process.

As mentioned in the previous sections, VLM is not good at showing/tracking swaps, additions, or deletions in the HTML. Therefore, without images, responses are better when the questions have swaps, insertions, and deletions of elements (TRACKING SHUFFLED OBJECTS). In DATE UNDERSTANDING and NAVIGATE tasks, the images generated from HTML often fail to accurately represent the questions. In DATE UNDERSTANDING, LLAVA-1.5 generated HTML can not fully maintain the date, month, and year pattern mentioned in the question text (Table 11) which further confuses the VLM while performing reasoning with the image. Similarly, in NAVIGATE, GEMINI PRO generated HTML can not effectively depict the progression of navigation steps mentioned in the question text.

Image helps a different subset of a particular task. We further investigate the performance of with-image and without-image setups to find out when having an image is beneficial and when having an image hurts. As shown in Figure 9, we break down the performance by tasks and count the number of times the VLM produces a correct answer with an image and gets the same question wrong without an image (*Image Improves*). Then, we reverse the conditions, i.e., count the number of times the VLM produces an incorrect answer with an image and generates the correct answer for the same question without an image (*Image Hurts*). We can see that for all tasks, having images helps solve some questions that can not be solved without images and vice versa.

6 RELATED WORKS

Visual Problem Solving. Visual problem-solving is an inherent human ability while performing complex multi-step reasoning. Humans tend to draw a mental image of a question to understand and ground the problem, which helps to plan subsequent steps to solve the question. Mental images provide a simplified representation of the content of the cognitive task. Thus, the involvement of visual images in problem-solving may result in a notable degree of success (Bauer & Johnson-Laird, 1993; Antonietti, 1991). Studies also show that an accurate visual representation of a problem enhances the chances of solving the problem rather than having no representation at all (van Garderen et al., 2018; Krawec, 2014).

Reasoning with LLMs and VLMs. In recent years, several LLMs and VLMs have been introduced, which are showing impressive performance in complex reasoning tasks (OpenAI, 2023; Touvron et al., 2023; Chowdhery et al., 2022; Liu et al., 2023a; Zhu et al., 2023; Li et al., 2023; Dai et al., 2023; Liu et al., 2023b). However, when it comes to solving math word problems (Cobbe et al., 2021; Koncel-Kedziorski et al., 2016; Patel et al., 2021) or symbolic reasoning tasks (Suzgun et al., 2022), the VLM can not fairly compete with the LLM as the nature of these tasks is unimodal. While considerable efforts have been invested in improving the performance of LLMs on these reasoning tasks during inference (Madaan et al., 2023; Wang et al., 2023; Gao et al., 2023b; Wei et al., 2023; Poesia et al., 2023; Hao et al., 2023), fewer endeavors have been made to tackle these challenges from the perspective of a vision-language model (Lee et al., 2023; Hsu et al., 2023). A very relevant work to ours is Hsu et al. (2023), which leverages LLM to generate drawing commands and reads out abstractions from the resulting picture. However, it relies on a fine-tuned visual foundation model (Lee et al., 2023) to interpret abstractions from the drawn diagram, requiring additional training data. In addition, diagrams can only benefit specific tasks, limiting their applicability to diverse reasoning types. In this paper, we study these text-only benchmarks using VLMs by proposing a simple idea to leverage the full potential of a VLM on diverse reasoning tasks.

7 CONCLUSION

In this work, we present SELF-IMAGINE, an approach that maximizes the capabilities of Vision Language Models (VLMs) in solving text-only reasoning tasks. Our method draws on a common human problem-solving technique, creating visual representations of problems to aid in reasoning. Our approach is self-sufficient, requiring no additional data, supervision, or training. Through our intensive experiments with diverse reasoning tasks, we find that SELF-IMAGINE significantly improves the performance of stae-of-the-art VLMS (LLAVA-1.5 & GEMINI PRO) using self-generated images. We also find that the extent of improvement relies heavily on the quality of the generated image. We present cases where image improves and hurts the performance of the VLM, motivating future research on better image generation approaches.

REFERENCES

Syeda Nahida Akter, Zichun Yu, Aashiq Muhamed, Tianyue Ou, Alex Bäuerle, Ángel Alexander Cabrera, Krish Dholakia, Chenyan Xiong, and Graham Neubig. An in-depth look at gemini's language abilities, 2023.

Alessandro Antonietti. Chapter 15 why does mental visualization facilitate problem-solving? In Robert H. Logie and Michel Denis (eds.), *Mental Images in Human Cognition*, volume 80 of *Advances in Psychology*, pp. 211–227. North-Holland, 1991. doi: https://doi.org/10.1016/S0166-4115(08)60515-4. URL `https://www.sciencedirect.com/science/article/pii/S0166411508605154`.

Malcolm I. Bauer and P. N. Johnson-Laird. How diagrams can improve reasoning. *Psychological Science*, 4(6):372–378, 1993. ISSN 09567976, 14679280. URL `http://www.jstor.org/stable/40062565`.

A.J.H. Boonen, F. van Wesel, J. Jolles, and M. van der Schoot. The role of visual representation type, spatial ability, and reading comprehension in word problem solving: An item-level analysis in elementary school children. *International Journal of Educational Research*, 68(4):15–26, 2014. ISSN 0883-0355. doi: 10.1016/j.ijer.2014.08.001.

Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. Visual instruction tuning with polite flamingo. *arXiv preprint arXiv:2307.01003*, 2023a.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechu Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023b.

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, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

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. Palm: Scaling language modeling with pathways, 2022.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023a. URL `https://zenodo.org/records/10256836`.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10764–10799. PMLR, 23–29 Jul 2023b. URL https://proceedings.mlr.press/v202/gao23f.html.

GoogleAI. Bard, 2023.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 8154–8173, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.507. URL https://aclanthology.org/2023.emnlp-main.507.

Joy Hsu, Gabriel Poesia, Jiajun Wu, and Noah D. Goodman. Can visual scratchpads with diagrammatic abstractions augment LLM reasoning? In *I Can't Believe It's Not Better Workshop: Failure Modes in the Age of Foundation Models*, 2023. URL https://openreview.net/forum?id=YlhKbQ0zF3.

Rik Koncel-Kedziorski, Subhro Roy, Aida Amini, Nate Kushman, and Hannaneh Hajishirzi. MAWPS: A math word problem repository. In Kevin Knight, Ani Nenkova, and Owen Rambow (eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1152–1157, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1136. URL https://aclanthology.org/N16-1136.

Jennifer L. Krawec. Problem representation and mathematical problem solving of students of varying math ability. *Journal of Learning Disabilities*, 47(2):103–115, 2014. doi: 10.1177/0022219412436976. URL https://doi.org/10.1177/0022219412436976. PMID: 22392891.

Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pp. 18893–18912. PMLR, 2023.

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 *ICML*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023a.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. Llava-plus: Learning to use tools for creating multimodal agents, 2023b.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Sean Welleck, Bodhisattwa Prasad Majumder, Shashank Gupta, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback, 2023.

Shen-yun Miao, Chao-Chun Liang, and Keh-Yih Su. A diverse corpus for evaluating and developing english math word problem solvers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 975–984, 2020.

OpenAI. Openai gpt-3.5 api [gpt-3.5-turbo], 2021.

OpenAI. Gpt-4 technical report, 2023.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168.

Gabriel Poesia, Kanishk Gandhi, Eric Zelikman, and Noah D. Goodman. Certified reasoning with language models. 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions, 2020.

Simon St.Laurent, Murata Makoto, and Dan Kohn. XML Media Types. RFC 3023, January 2001. URL https://www.rfc-editor.org/info/rfc3023.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

Till Tantau. *The TikZ and PGF Packages*, April 2022. URL https://github.com/pgf-tikz/pgf. Manual for version 3.1.9a.

Gemini Team. Gemini: A family of highly capable multimodal models, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Delinda van Garderen, Amy Scheuermann, and Christa Jackson. Developing representational ability in mathematics for students with learning disabilities: A content analysis of grades 6 and 7 textbooks. *Learning Disability Quarterly*, 35(1):24–38, 2012. doi: 10.1177/0731948711429726. URL https://doi.org/10.1177/0731948711429726.

Delinda van Garderen, Amy Scheuermann, Apryl Poch, and Mary M. Murray. Visual representation in mathematics: Special education teachers' knowledge and emphasis for instruction. *Teacher Education and Special Education*, 41(1):7–23, 2018. doi: 10.1177/0888406416665448. URL https://doi.org/10.1177/0888406416665448.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chaoya Jiang, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest, 2023.

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

A APPENDIX

B PROMPTS

The prompt used for image generation is shown in Listing 1. Please see the code repository for the complete prompt.

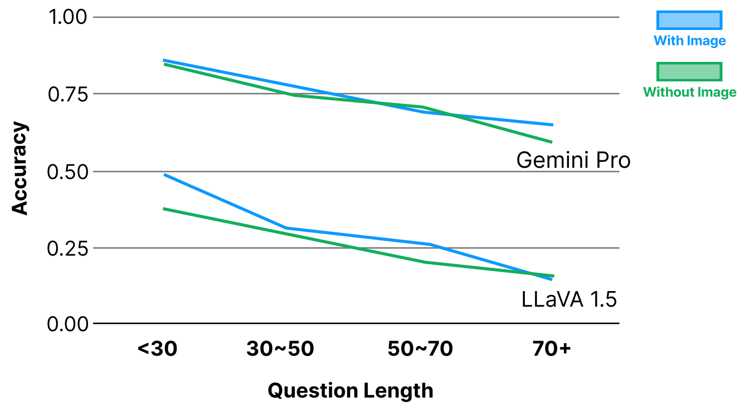
```

Q: Alfie, the albatross, flies 400 kilometers every day.  If the
    circumference of the earth is 40,000 kilometers, how many days will
    it take Alfie to fly a distance equal to half of the way around the
    earth?

# HTML code:


<!DOCTYPE html>
<html lang="en">
<head>
<meta charset="UTF-8">
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title>Alfie’s Journey</title>
<style>
  .diagram-container {{
    display: flex;
    align-items: center;
    justify-content: center;
    flex-direction: column;
    font-family: Arial, sans-serif;
  }}

  .earth {{
    position: relative;
    width: 200px;
    height: 200px;
    border: 3px solid blue;
    border-radius: 50%;
    overflow: hidden;
  }}



```

```
  .text {{
    margin: 10px;
    text-align: center;
  }}

  .stat {{
    display: flex;
    justify-content: space-around;
    margin-top: 20px;
  }}

  .stat > div {{
    text-align: center;
  }}
</style>
</head>
<body>

<div class="diagram-container">
  <div class="earth">
    <div class="albatross-flight"></div>
  </div>
  <div class="text">Alfie's Journey Around the Earth</div>
  <div class="stat">
    <div>
      <strong>Alfie's Daily Distance:</strong><br>
      400 km
    </div>
    <div>
      <strong>Earth's Circumference:</strong><br>
      40,000 km
    </div>
    <div>
      <strong>Target Distance:</strong><br>
      20,000 km (halfway around the Earth)
    </div>
  </div>
</div>

</body>
</html>
```

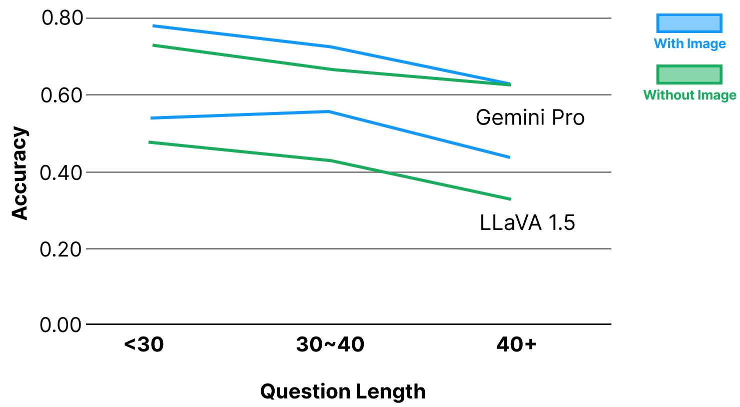
Listing 1: Prompt for generating HTML using VLM

Task	Prompt	
	Question Only	Question + Image
GSM8K, ASDIV, SVAMP, MAWPS	Solve the math problem. Think step-by-step. Always end your answer with 'The answer is ⟨answer⟩'.	Solve the math problem using the image. Think step-by-step. Always end your answer with 'The answer is ⟨answer⟩'.
PENGUINS IN A TABLE	Answer questions about a table of penguins and their attributes.	Answer questions about a table of penguins and their attributes using the image.
COLORED OBJECTS	Answer extremely simple questions about the colors of objects on a surface.	Answer extremely simple questions about the colors of objects on a surface using the image.
OBJECT COUNTING	Questions that involve enumerating objects and asking the model to count them.	Questions that involve enumerating objects and asking the model to count them using the image.
NAVIGATE	Given a series of navigation instructions, determine whether one would end up back at the starting point.	Given a series of navigation instructions, determine whether one would end up back at the starting point using the image.
DATE UNDERSTANDING	Infer the date from context.	Infer the date from context using the image.
GEOMETRIC SHAPES	Name geometric shapes from their SVG paths.	Name geometric shapes from their SVG paths and using the image.
TEMPORAL SEQUENCES	Answer questions about which times certain events could have occurred.	Answer questions about which times certain events could have occurred using the image.
TRACKING SHUFFLED OBJECTS	A task requiring determining the final positions of a set of objects given their initial positions and a description of a sequence of swaps.	A task requiring determining the final positions of a set of objects given their initial positions and a description of a sequence of swaps using the image.

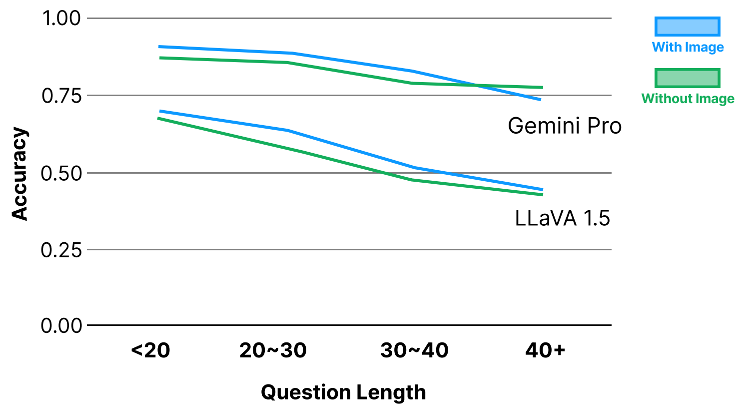
Table 3: Prompts used for both reasoning and mathematics tasks. For all reasoning tasks, we also add ***Please think step-by-step, and finally answer by selecting an option using the format "The answer is ⟨option⟩"*** after adding the question to the above mentioned prompts.



(a) GSM8K



(b) SVAMP



(c) ASDIV

Figure 6: Accuracy by question length across three mathematical reasoning tasks. In the cases of ASDIV and SVAMP, accuracy is notably higher when utilizing images for longer and more intricate questions compared to scenarios without images. However, in the context of more complex questions, such as those found in GSM8K, the limitations of the VLM become apparent. In this scenario, the inability to generate effective HTML results in erroneous image generation, consequently leading to decreased accuracy, particularly with longer questions.

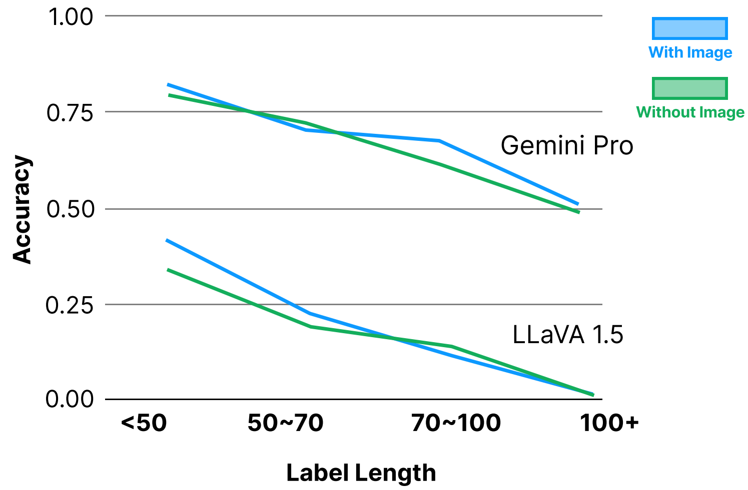


Figure 7: GSM8K accuracy by chain-of-thought length. Similar to the findings in Figure 6, image representations for complex questions are not efficient and structured. Therefore, the inclusion of images does not enhance the representation of questions that demand longer chains of thought.

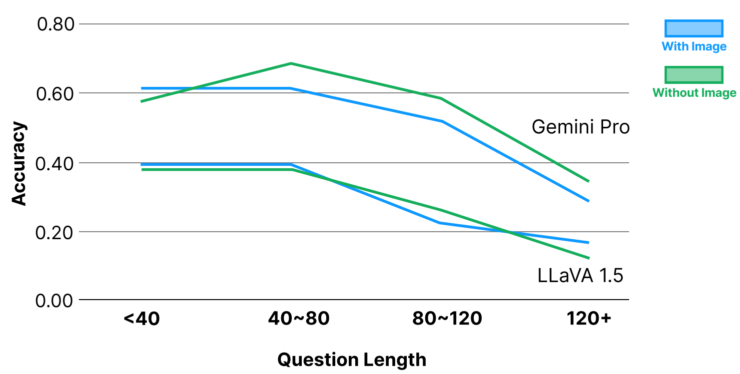


Figure 8: Accuracy by Question Length for a subset of BIG-Bench-Hard benchmark. Incorporating images helps when the corresponding question is simpler and shorter and when the questions are more complex.

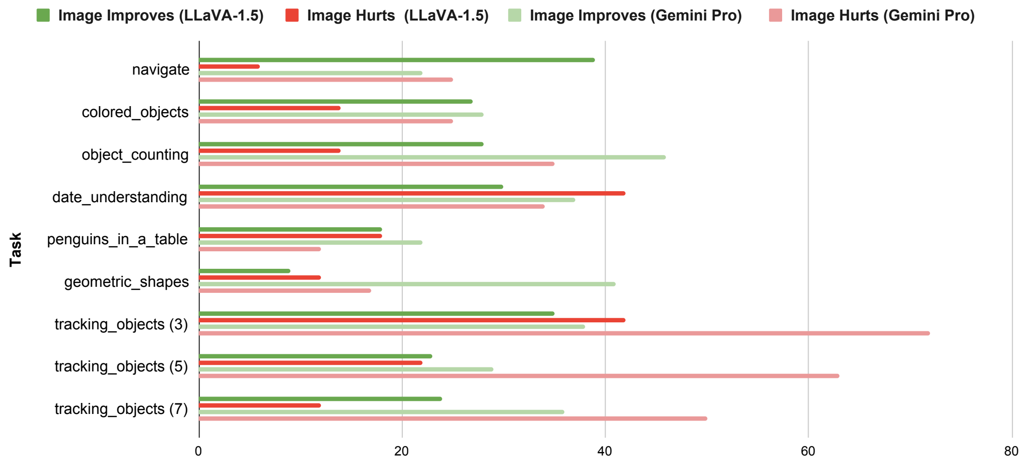


Figure 9: Number of Instances from each subtask impacted by Image. Here *'Image Hurts'* represents instances that achieved correct answers without image and got incorrect with image. Similarly *'Image Improves'* shows data points getting the correct answers with image and getting incorrect without image.

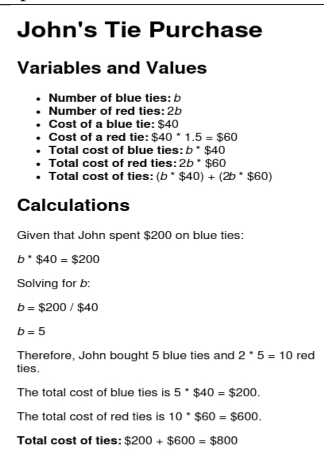
Question	John buys twice as many red ties as blue ties. The red ties cost 50% more than blue ties. He spent \$200 on blue ties that cost \$40 each. How much did he spend on ties?
Image	 **John's Tie Purchase** **Variables and Values** • **Number of blue ties:** *b* • **Number of red ties:** 2*b* • **Cost of a blue tie:** \$40 • **Cost of a red tie:** \$40 * 1.5 = \$60 • **Total cost of blue ties:** *b* * \$40 • **Total cost of red ties:** 2*b* * \$60 • **Total cost of ties:** (*b* * \$40) + (2*b* * \$60) **Calculations** Given that John spent \$200 on blue ties: *b* * \$40 = \$200 Solving for *b*: *b* = \$200 / \$40 *b* = 5 Therefore, John bought 5 blue ties and 2 * 5 = 10 red ties. The total cost of blue ties is 5 * \$40 = \$200. The total cost of red ties is 10 * \$60 = \$600. **Total cost of ties:** \$200 + \$600 = \$800
Label	800
Output With Image	Let the number of blue ties be b. The number of red ties is twice the number of blue ties, so there are $2b$ red ties. The cost of a blue tie is \$40, and the cost of a red tie is 50% more than the cost of a blue tie, so the cost of a red tie is $40 * 1.5 = \$60$. The total cost of the blue ties is $b * 40$, and the total cost of the red ties is $2b * 60$. The total cost of the ties is $b * 40 + 2b * 60$. We know that John spent \$200 on blue ties, so we can write the equation $b * 40 = 200$. Solving for b, we get $b = 5$. Therefore, John bought 5 blue ties and $2 * 5 = 10$ red ties. The total cost of the blue ties is $5 * 40 = \$200$. The total cost of the red ties is $10 * 60 = \$600$. The total cost of the ties is $200 + 600 = \$800$. The answer is **\$800**.
Output Without Image	John bought 5 blue ties and 10 red ties. The blue ties cost \$200, so the red ties cost \$300. He spent a total of \$500 on ties. The answer is **500**.

Table 4: Example of Image improving reasoning in GSM8K task for GEMINI PRO.

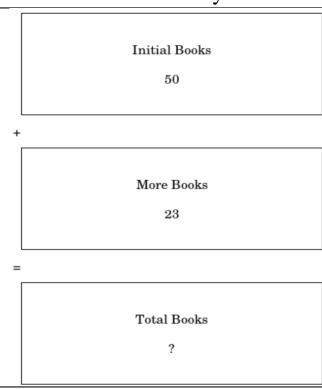
Question	Our class got fifty-four books from the library. Then we got twenty-three more books from the library. How many books did our class get from the library?
Image	
Label	77
Output With Image	To solve this problem, we need to add the initial number of books (50) and the number of books received from the library (23). 50 + 23 = 73 The answer is **73**.
Output Without Image	To solve this problem, we need to add the number of books the class got from the library initially (54) to the number of books they got from the library later (23). Step 1: Add 54 and 23 to find the total number of books the class got from the library. 54 + 23 = 77 The answer is **77**.

Table 5: Example of Image hurting reasoning in ASDIV task for LLAVA-1.5.

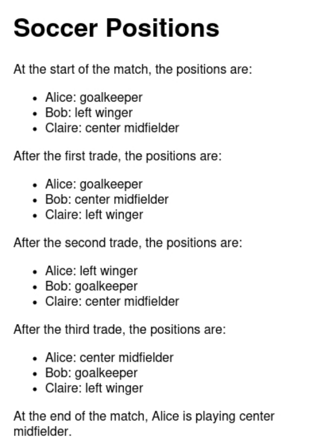
Question	Alice, Bob, and Claire are on the same team in a soccer match. At the start of the match, they are each assigned to a position: Alice is playing goalkeeper, Bob is playing left winger, and Claire is playing center midfielder. As the game progresses, pairs of players occasionally swap positions. First, Claire and Bob trade positions. Then, Alice and Bob trade positions. Finally, Claire and Alice trade positions. At the end of the match, Alice is playing Options: (A) goalkeeper (B) left winger (C) center midfielder
Image	 **Soccer Positions** At the start of the match, the positions are: • Alice: goalkeeper • Bob: left winger • Claire: center midfielder After the first trade, the positions are: • Alice: goalkeeper • Bob: center midfielder • Claire: left winger After the second trade, the positions are: • Alice: left winger • Bob: goalkeeper • Claire: center midfielder After the third trade, the positions are: • Alice: center midfielder • Bob: goalkeeper • Claire: left winger At the end of the match, Alice is playing center midfielder.
Label	(B)
Output With Image	After the first trade, Alice is still the goalkeeper, Bob is now the center midfielder, and Claire is the left winger. After the second trade, Alice is now the left winger, Bob is the goalkeeper, and Claire is still the center midfielder. After the third trade, Alice is the center midfielder, Bob is the left winger, and Claire is the goalkeeper. At the end of the match, Alice is playing center midfielder. The answer is **(C)**.
Output Without Image	First, Claire and Bob trade positions. This means that Claire is now playing left winger and Bob is playing center midfielder. Next, Alice and Bob trade positions. This means that Alice is now playing center midfielder and Bob is playing goalkeeper. Finally, Claire and Alice trade positions. This means that Claire is now playing goalkeeper and Alice is playing left winger. Therefore, at the end of the match, Alice is playing left winger. The answer is **(B)**.

Table 6: Example of Image hurting reasoning in TRACKING SHUFFLED OBJECTS of three objects task for GEMINI PRO.

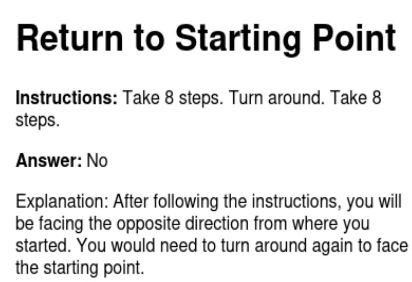
Question	If you follow these instructions, do you return to the starting point? Take 8 steps. Turn around. Take 8 steps. Options: - Yes - No
Image	
Label	(B)
Output With Image	The answer is **No**. You start by taking 8 steps forward. Then you turn around, which means you are now facing the opposite direction. Then you take 8 steps forward again. This means you are now 8 steps away from your starting point, but facing the opposite direction. You would need to turn around again to face the starting point.
Output Without Image	The answer is **yes**.

Table 7: Example of Image hurting reasoning in NAVIGATE task for GEMINI PRO.

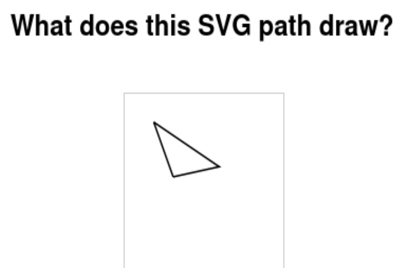
Question	This SVG path element ⟨path d="M 30.17,45.97 L 58.79,40.36 L 18.10,15.70 M 18.10,15.70 L 30.17,45.97"/⟩draws a Options: (A) circle (B) heptagon (C) hexagon (D) kite (E) line (F) octagon (G) pentagon (H) rectangle (I) sector (J) triangle
Image	
Label	(J)
Output With Image	The path starts at point (30.17,45.97) and draws a line to point (58.79,40.36). Then it draws another line to point (18.10,15.70). Finally, it draws a line back to the starting point (30.17,45.97). The resulting shape is a triangle. The answer is **(J)**.
Output Without Image	The path element consists of two subpaths. The first subpath starts at (30.17, 45.97) and goes to (58.79, 40.36), then (18.10, 15.70). The second subpath starts at (18.10, 15.70) and goes back to (30.17, 45.97). The first subpath looks like a line from (30.17, 45.97) to (18.10, 15.70), and the second subpath looks like a line from (18.10, 15.70) to (30.17, 45.97). So it should be a line. The answer is **(E)**.

Table 8: Example of Image improving reasoning in GEOMETRIC SHAPES task for GEMINI PRO.

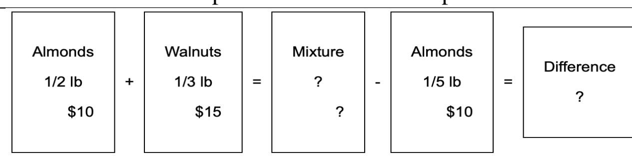
Question	A pound of almonds costs $10 while a pound of walnuts costs $15. How much more does it cost for a mixture of 1/2 pound almonds and 1/3 pound walnuts than a mixture of 1/5 pound almonds and 1/3 pound walnuts?
Image	
Label	3
Output With Image	To solve this problem, we need to find the cost of each mixture and then compare them. Cost of 1/2 pound almonds and 1/3 pound walnuts: We know that a pound of almonds costs $10, and a pound of walnuts costs $15. To find the cost of 1/2 pound almonds and 1/3 pound walnuts, we need to multiply the cost per pound by the amount of each nut in the mixture: (10/lb) * (1/2 lb almonds) + (15/lb) * (1/3 lb walnuts) = $10 + $5 = $15 Cost of 1/5 pound almonds and 1/3 pound walnuts: We can find the cost of this mixture in a similar way: (10/lb) * (1/5 lb almonds) + (15/lb) * (1/3 lb walnuts) = $2 + $5 = $7 Comparison: Now we can compare the costs of the two mixtures: $15 - $7 = $8 The answer is **$8**.
Output Without Image	To solve this problem, we need to find the cost of the mixture of 1/2 pound almonds and 1/3 pound walnuts and compare it to the cost of the mixture of 1/5 pound almonds and 1/3 pound walnuts. First, we need to find the cost of 1/2 pound of almonds and 1/3 pound of walnuts. We can do this by multiplying the cost per pound of almonds and walnuts by the respective weights: 1/2 pound of almonds * $10/pound = $1/2*10=5 1/3 pound of walnuts * $15/pound = $1/3*15=5 The total cost of the mixture is the sum of the costs of the almonds and walnuts: $5 + $5 = $5+5=10 Next, we need to find the cost of 1/5 pound of almonds and 1/3 pound of walnuts. We can do this in the same way: 1/5 pound of almonds * $10/pound = $1/5*10=2 1/3 pound of walnuts * $15/pound = $1/3*15=5 The total cost of the mixture is the sum of the costs of the almonds and walnuts: $2 + $5 = $2+5=7 Now we can compare the costs of the two mixtures: $10 - $7 = $10-7=3 The answer is **$3**.

Table 9: Example of Image hurting reasoning in GSM8K task for LLAVA-1.5.

Question	If you follow these instructions, do you return to the starting point? Always face forward. Take 5 steps forward. Take 8 steps backward. Take 4 steps forward. Take 4 steps right. Options: - Yes - No
Image	No, you will not return to the starting point if you follow these instructions. The sequence of steps is as follows: 1. 7 steps left: 7 steps to the left 2. 2 steps backward: 2 steps backward 3. 7 steps backward: 7 steps backward 4. 7 steps backward: 7 steps backward 5. 3 steps forward: 3 steps forward After completing these steps, you will be 7 steps to the left of your starting position, and then you will take 2 steps backward, which will put you 2 steps behind your starting position. You will then take 7 steps backward, which will put you 7 steps further behind your starting position. Finally, you will take 3 steps forward, which will bring you 3 steps closer to your starting position. However, you will still be 4 steps behind your starting position (7 steps left + 2 steps backward + 7 steps backward + 7 steps backward - 3 steps forward = 4 steps behind).
Label	No
Output With Image	The answer is **No**
Output Without Image	To determine whether one would end up back at the starting point, we need to analyze the given navigation instructions step-by-step. Take 5 steps forward. Take 8 steps backward. Take 4 steps forward. Take 4 steps right. Let's analyze each step: Take 5 steps forward: After taking 5 steps forward, the person would be 5 steps away from the starting point. Take 8 steps backward: After taking 8 steps backward, the person would be 8 steps away from the starting point, in the opposite direction. Take 4 steps forward: After taking 4 steps forward, the person would be 4 steps away from the starting point, but now they are moving in the correct direction. Take 4 steps right: After taking 4 steps right, the person would be 4 steps away from the starting point, but now they are at a right angle to the starting point. Since the person is now at a right angle to the starting point and moving in the correct direction, they will eventually return to the starting point by continuing to move forward. The answer is: **Yes**

Table 10: Example of Image improving reasoning in NAVIGATE task for LLAVA-1.5.

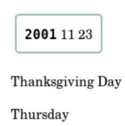
Question	TIn the US, Thanksgiving is on the fourth Thursday of November. Today is the US Thanksgiving of 2001. What is the date yesterday in MM/DD/YYYY? Options: (A) 09/12/2002 (B) 11/30/2002 (C) 11/21/2002 (D) 11/21/2076 (E) 11/07/2002 (F) 11/15/2002
Image	
Label	(C)
Output With Image	The answer is **(E)** 11/07/2002.
Output Without Image	To infer the date from context, we need to determine the date of US Thanksgiving in 2001. Step 1: Determine the date of US Thanksgiving in 2001. The US Thanksgiving holiday is celebrated on the fourth Thursday of November. In 2001, November has 30 days. Step 2: Calculate the date of US Thanksgiving in 2001. 4th Thursday of November 2001: Thursday = 22 (since it's a Thursday) November = 30 22 + 30 = 52 (the date in numerical format) Step 3: Convert the numerical date to MM/DD/YYYY format. 52 = 11/22/2001 The answer is **(C)** 11/21/2002.

Table 11: Example of Image hurting reasoning in DATE UNDERSTANDING task for LLAVA-1.5.