Chain of Images for Intuitively Reasoning

Fanxu Meng[1,2], **Haotong Yang**[1], **Yiding Wang**[1], **Muhan Zhang**[1,2]
[1]School of Intelligence Science and Technology, Peking University
[2]National Key Laboratory of General Artificial Intelligence, BIGAI
https://github.com/GraphPKU/CoI

Abstract

The human brain is naturally equipped to comprehend and interpret visual information rapidly. When confronted with complex problems or concepts, we use flowcharts, sketches, and diagrams to aid our thought process. Leveraging this inherent ability can significantly enhance logical reasoning. However, current Large Language Models (LLMs) do not utilize such visual intuition to help their thinking. Even the most advanced vision language models (e.g., GPT-4V and LLaVA) merely align images into the textual space, which means their reasoning processes remain purely verbal. To mitigate such limitations, we present a Chain of Images (CoI) approach, which can convert complex language reasoning problems to simple pattern recognition by generating a series of images as intermediate representations. Furthermore, we have developed a CoI evaluation dataset encompassing 15 distinct domains where images can intuitively aid problem-solving. Based on this dataset, we aim to construct a benchmark to assess the capability of future multimodal LLMs to leverage images for reasoning. In supporting our CoI reasoning, we introduce a symbolic multimodal large language model (SyMLLM) that generates images strictly based on language instructions and accepts both text and image as input. Experiments on Geometry, Chess and Common Sense tasks sourced from the CoI evaluation dataset show that CoI improves performance significantly over the pure-language Chain of Thoughts (CoT) baselines.

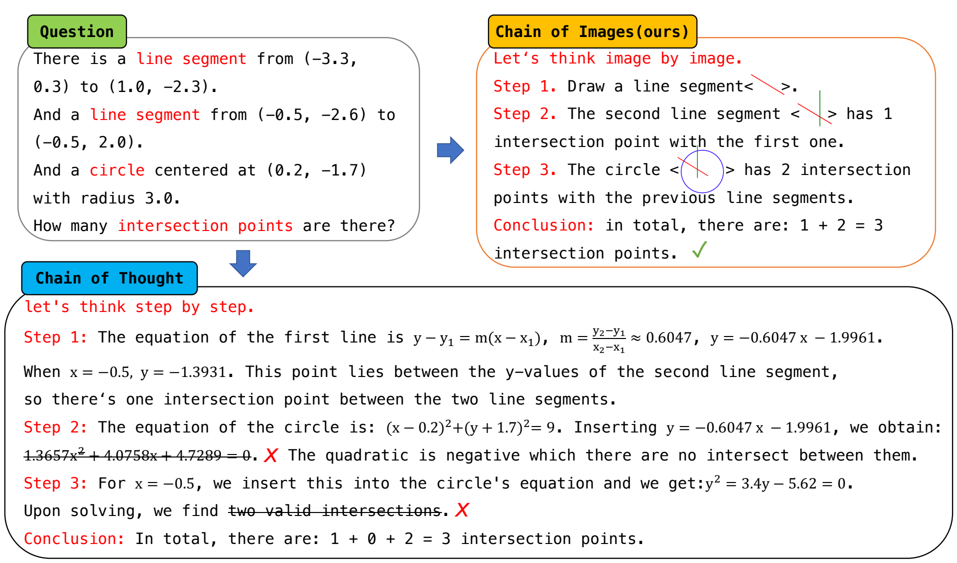


Figure 1: The simplified steps employing GPT-4 to calculate the number of intersection points using CoT are displayed in the bottom block. The whole processes of CoI are displayed in the top right block. Two issues were identified in CoT: 1) Incorrect numerical values were used during the formula derivation, and 2) The existence of endpoints in the line segments was overlooked. By contrast, CoI easily identifies the number of intersection points in the image generated by SyMLLM.

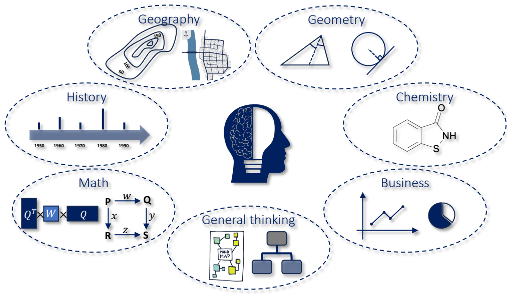


Figure 2: Images play a pivotal role in many disciplines. We tend to imagine pictures to solve problems intuitively.

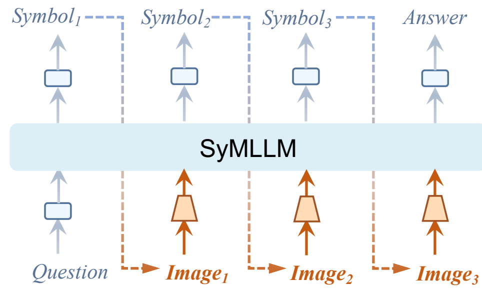


Figure 3: The symbolic response generated by SyMLLM given language instruction can be directly converted to an image.

1 Introduction

Large-scale pre-trained language models (LLMs) (Vaswani et al., 2017; Devlin et al., 2018; Brown et al., 2020; Chowdhery et al., 2022; Touvron et al., 2023; OpenAI, 2023) have found applications in diverse areas. Beyond their robust foundational capabilities (Kaplan et al., 2020; Ouyang et al., 2022), a significant advancement lies in the evolution of prompting techniques (Liu et al., 2023b), which provide LLMs with a context to guide their responses and reasoning processes. As presented by Brown et al. (2020), the few-shot prompting approach offers the model several question-answer pairs to guide its outputs. The Chain-of-Thought (CoT) prompting strategy (Wei et al., 2022) provides sequential, step-by-step answer examples, enabling the model to perform intricate multi-step reasoning. The zero-shot CoT technique (Kojima et al., 2022) omits to provide examples, instead stimulating the model to think progressively with a simple "Let's think step by step" prompt. For instance, in Figure 1, to accurately count the intersection points of geometric shapes, CoT follows a sequential approach: 1) Determine the formula of each shape, 2) Compute the intersection points between each pair of shapes, and 3) Sum up all the intersection points.

However, the language-based reasoning process is often overly complex and abstract. Using images as a medium to solve textual logical problems is very intuitive. Humans excel at recalling visual concepts. For instance, one can easily recall someone's appearance while swiftly forgetting their name. The human brain is inherently adept at intuitive reasoning. We constantly employ imagery to facilitate thought processes in our daily lives. And many academic disciplines, as illustrated in Figure 2, utilize images to aid understanding and comprehension.

Drawing inspiration from the human brain's proficiency in intuitive reasoning, we have developed the Chain-of-Images (CoI) method to generate images to help solve complex and abstract problems. As illustrated in the right top of Figure 1, we generate images for the three shapes step by step. Starting from the second step, we identify the intersection points between the newly generated shape and the shapes from previous steps in the image. While algorithmically computing intersection points is challenging for LLMs, with the images, discerning the count of intersection points is very easy for an image encoder with good pattern recognition capabilities.

As a preliminary attempt, we point out that our Chain of Images method broadly offers at least but far more than these three advantages:

- Images provide an intuitive representation of the relationships (spatial, topological, temporal, etc.) between the items for tasks requiring complex relational reasoning, such as counting the intersection points of geometric shapes.
- Images assist in providing a compact representation that can record the current state at each step for tasks that satisfy Markovian properties and involve extended reasoning chains, like playing chess.
- Images capture a wealth of commonsense knowledge not present in language alone. Incorporating images into the reasoning process enhances its robustness.

In supporting our CoI reasoning, we introduce a symbolic multimodal large language model (SyMLLM) shown in Figure 3 that generates images strictly based on language instructions and accepts both text and image as input. Furthermore, to validate the value of CoI, we have developed a CoI evaluation dataset, including 15 distinct tasks. This dataset is intended to assess how images can facilitate reasoning quantitatively.

Empirical results on geometric, chess, and commonsense tasks from the CoI evaluation dataset demonstrate that the SyMLLM framework can generate the required images with nearly 100% accuracy. Based on such accurate image generation, SyMLLM employing CoI significantly outperforms the same-structure LLM that relies on pure text for reasoning on these tasks. For instance, the accuracy of calculating the number of intersection points among four shapes increased from 27.75% to 64.25%.

2 Chain of Images Evaluation Dataset

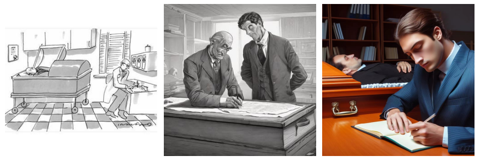
Current MLLMs (Wu et al., 2023; Dong et al., 2023; Zheng et al., 2023a) have made significant progress in generating high-quality image details and understanding images. Yet, in many tasks, a graphical imagination is as important as or even more important to humans as an intermediate step in reasoning than a sequential one. The next frontier for these models is evaluating and enhancing their capacity for integrated reasoning with text and images. To evaluate this capability, we introduced the CoI evaluation dataset, or CoIEval, which measures explicitly MLLMs' proficiency in utilizing image generation as a step in textual reasoning processes.

We employ GPT-4 to filter through various existing evaluation datasets, including BIGBench (Srivastava et al., 2022), AGIEval (Zhong et al., 2023), MMLU (Hendrycks et al., 2020), NYCCHessel et al. (2023) to isolate questions suitable for CoI tasks. To guide GPT-4 in recognizing when image generation can be beneficial for problem-solving, we designed a specialized 3-shot prompt. This prompt contains three distinct questions, depicted in Figure 4. It serves as an instruction for GPT-4 to identify where visual aids can enhance understanding and problem resolution. For example, problems in chess and geometry demonstrate how image generation can facilitate the comprehension of complex spatial relationships, whereas a simple arithmetic problem exemplifies a case where such visual support is unnecessary. The "QUESTION" should be replaced by the questions from each task.

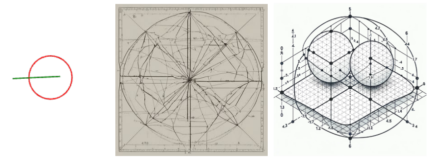
```
Q: Is it more advantageous to address these tasks by generating images for a more intuitive solution? Why?
"In the following chess position, find a checkmate-in-one move. 1. d4 g6 2. c3 d5 3. Qd3 Bf5 4. Qe3 Nc6 5. h3 Nf6 6. g4 Be6
7. f3 Bg7 8. Nd2 Qd6 9. Kf2 h5 10. Qd3 hxg4 11. hxg4 Rxh1 12. Kg2 Rh4 13. Nh3 Bxg4 14. fxg4 Rxg4+ 15. Kh1 O-O-O 16. Nf3 Rh8."
A: Yes, it's often easier for many people to understand a chess problem by visualizing the position on a board rather than
parsing through a series of moves in descriptive notation. Generating an image of the board position can help a player
quickly identify threats, patterns, and potential moves.
Q: Is it more advantageous to address these tasks by generating images for a more intuitive solution? Why?
"There is a line segment from (-3.8, -1.1) to (1.7, -2.1). There is a polygon with coordinates [(-1.2, 3.6), (0.8, -3.2),
(4.6, 2.2)]. There is a circle centered at (-2.0, -0.2) with radius 3.0. How many intersection points are there?"
A: Yes, visual representation is definitely useful for problems like this, as it can quickly give a sense of the relative
positioning and potential intersections of the various geometrical shapes.
Q: Is it more advantageous to address these tasks by generating images for a more intuitive solution? Why?
"What is 4 plus 2?"
A: No, this problem is straightforward and can be easily solved without the aid of visual tools.
Q: Is it more advantageous to address these tasks by generating images for a more intuitive solution? Why?
"QUESTION"
A:
```

Figure 4: the 3-shot prompt for building the CoIEval dataset.

Based on the 3-shot prompt, we get a collection of 15 tasks across various domains, as shown in Appendix 4. In this table, each row is divided into three lines: Line 1. The task name and the source dataset from which the task is derived. Line 2. An example question from the task. Line 3. GPT-4's response regarding whether image generation aids in reasoning—indicating 'Yes' or 'No', along with the rationale for why GPT-4 believes an image would or would not be helpful.



(a) “A man at a desk writing something down. Another man is lying behind him in a coffin.” “SDXL fail at processing the location relation between two men and a coffin.”



(b) There is a circle centered at (1.6, -0.2) with radius 2.0. There is a line segment from (2.5, -0.3) to (-1.9, -0.1). How many intersection points are there? “SDXL and DALL·E 3 unable to generate simple geometric shapes.”

Figure 5: This figure demonstrates the generation of new images by SDXL (center) and DALL·E 3 (right), using the captions derived from the original images (left).

3 Symbolic Multimodal Large Language Models

The primary step of CoI is to generate images step by step. Stable Diffusion XL (Podell et al., 2023) are state-of-the-art open-source image generative models which good at creating images with vast details. However, as illustrated in Figure 5a, we encountered a notable limitation of the SDXL model when it was tasked with following complex instructions to generate images. Additionally, in Figure 5b, SDXL could not draw simple shapes like a circle and a line segment. The first limitation could potentially be addressed by DALL·E 3, an advanced close-source image generation model from OpenAI. However, even DALL·E 3 encounters difficulties when generating abstract sketches or symbolic diagrams that require adherence to strict rules or relational constraints, as also evidenced in Figure 5b. Without an appropriate image, CoI cannot enhance the model’s reasoning capabilities because introducing noisy and disturbing images could, on the contrary, worsen the performance. Therefore, we call for an accurate and controllable image generation strategy. Based on this request, we introduce the Symbolic Multimodal LLM (SyMLLM), which is shown in Figure 3. It consists of an LLM, a symbol-to-image decoder, and an image encoder. When various text prompts are provided to the LLM, it will produce symbolic representations in different formats (e.g., SVG format), which can be directly transformed into bitmap images. Subsequently, these images are converted into image embeddings by the image encoder. The embeddings are then concatenated with the text embeddings to predict the next token. SyMLLM leverages the robust capabilities of LLMs to generate symbols precisely from language context and losslessly converts these symbols into image formats. In subsequent experiments, we observe that the accuracy of image generation approaches nearly 100%.

The symbol we use in this paper is Scalable Vector Graphics (SVG). SVG is a widely adopted XML-based format for creating two-dimensional graphics. Unlike bitmap graphics which are made up of pixels, the raw file of SVG graphics are composed of language that can directly generated by LLMs. Meanwhile, they can be shown as pixel images scaled up or down without any loss in quality. SVG offers the capability to produce graphics that range from simple shapes and icons (e.g. geometric shapes)to highly intricate and detailed illustrations (e.g. chess board and pieces). With the help of the SVG format, we can generate images only using the text-based LLM.

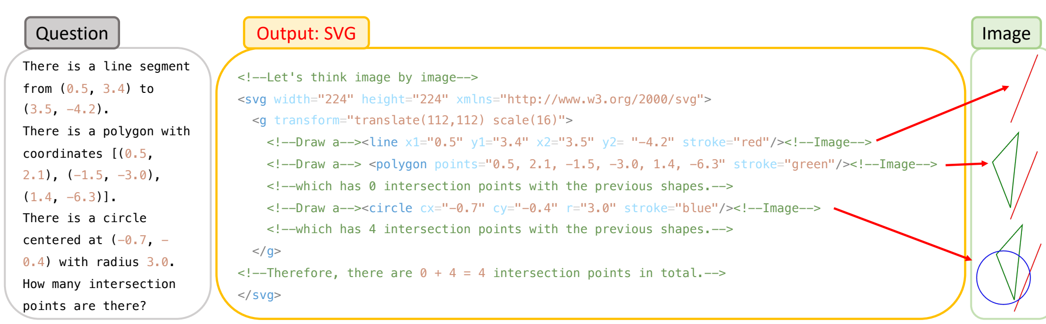


Figure 6: SVG format basic geometric shapes.

Figure 7: SVG format chess board and pieces.

As shown in Figure 6, for the problem of counting the intersections of geometric shapes, the output of SyMLLM is a vector image in SVG format. The instructions are written as code comment in the vector image: $<!--Let's\ think\ by\ image-->$, $<!--Draw\ a-->$, $<!--Image-->$, $<!--which\ has\ 4\ intersection\ points\ with\ the\ previous\ shapes.-->$, $<!--Therefore,\ there\ are\ 0+4=4\ intersection\ points\ in\ total.-->$. These instructions can be generated by the large model while do not affect the content of the pixel image converted. When come to $<!--Image-->$ tokens, we convert the vector images into pixel images and then input them into SyMLLM's image encoder. Based on the images, it is easy to compute the intersection points of the geometric shapes.

As shown in Figure 7, for the problem of predicting checkmate in one move, SVG first defines classes for different chess pieces using polygons, then draws an 8×8 chessboard, and finally moves each piece to its final position. Since the first two steps are fixed for any chess position, we can provide them as prompts to the large model and only need to generate the final position of each piece. When $<!--Image-->$ tokens appear, we convert the current state of the chessboard into a pixel image, which is then input into SyMLLM's image encoder. Based on the current state, it is easy to predict whether a checkmate is possible in the next move.

4 EXPERIMENTS

In this section, we have chosen three specific tasks from our CoIEval for evaluating the benefits of the CoI approach: Geometry, Chess, and Commensense reasoning tasks.

4.1 GEOMETRIC

Task Description: We conducted a comprehensive evaluation of the CoI approach to determine its efficacy in solving geometric problems. The evaluation was based on the Intersect Geometric task from the CoIEval. This task is designed to assess the ability of models to identify intersections among geometric shapes like line segments, circles and polygons. It contains a total of 250,000 examples, which are stratified into five levels of difficulty. The complexity is determined by the number of geometric shapes involved in each problem, which ranges from 2 to 6.

Data Processing: For the purpose of evaluation, we constructed a balanced set by randomly selecting 1,000 examples from each level of difficulty to create an evaluation set. Additionally, we generated a training set by selecting 10,000 examples from each difficulty level, ensuring that there was no overlap between the sets due to the method of non-replacement sampling. In total, the training set comprised 50,000 examples and was converted into a format outlined in Figure 6, whereas the evaluation set remained in its original form.

Model Structure: For our experiments, we utilized the Vicuna-7B-v1.5 (Zheng et al., 2023b), an enhanced derivative of Llama 2 (Touvron et al., 2023), pre-trained on user-shared conversations collected from ShareGPT[1]. The image encoder is the clip-vit-large-patch14 (Radford et al., 2021).

[1]https://sharegpt.com

Table 1: The intersect geometric task. The first row lists the number of geometric shapes in each sample, representing the different difficulty levels; The second and third rows compare the performance w/ or w/o image training. The bottom row reports the similarity between the image generated by SyMLLM and the ground truth.

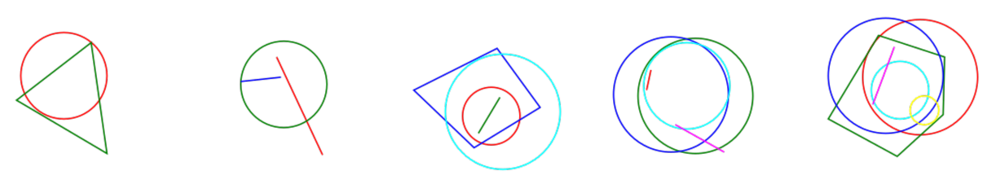
Number of Geometry Shapes	2	3	4	5	6
Text Accurcy	90.75	48.25	27.75	23.25	16.0
CoI Accurcy	**95.5**	**85.75**	**64.25**	**47.75**	**33.25**
Image Similarity	100	100	99.99	99.99	99.99

To establish a baseline for comparison, we trained a text-only model with identical hyper-parameters and prompts, excluding images during both training and inference.

Training Details: We following Dettmers et al. (2023) applied 4-bit quantization to the LLM's backbone and used a rank-16 LoRA, focusing on training newly added parameters. The fine-tuning process encompassed both the ViT and LLM, utilizing a single A800 GPU, with a batch size of 2 and 8 gradient accumulation steps, spanning over 5 epochs and completed within 12 hours.

Experiment Results: The results are listed in Table 1. From the table, we can see that the SyMLLM framework achieves nearly 100% accuracy in converting coordinates to an image. Moreover, the framework's ability to count intersection points within these images is significantly more effective compared to the baseline that relies solely on text. Specifically, the accuracy rate for detecting 4 intersection points has improved to $2.3\times$, while the detection of 5 intersection points shows an improvement of $2.05\times$.

Results Analysis: We have randomly selected one failure case from each level of difficulty and displayed it in Figure 8. It is evident that these failure cases are challenging to discern, even for the human eyes. One advantage of SVG lies in its scalability; it can be enlarged indefinitely without loss of quality. It is reasonable to believe that converting these graphics into larger images could further improve the performance of the CoI method on geometric intersection problems.



(a) $2 \rightarrow 4$ (b) $2 \rightarrow 1$ (c) $4 \rightarrow 6$ (d) $7 \rightarrow 9$ (e) $14 \rightarrow 18$

Figure 8: Failure cases for 2 to 6 geometric shapes, respectively. The number preceding the arrow indicates the ground truth number of intersection points, while the number following the arrow shows the count by the SyMLLM.

4.2 Chess

Task Description: We assess the effectiveness of CoI on the "Checkmate in One Move" task from the CoIEval benchmark. This task challenges models to identify a single move within a chess position that would result in a checkmate. In the context of chess, "checkmate" signifies a condition where the king is under immediate threat of capture with no legal moves available to escape. This task's complexity stems from the intricate series of moves leading up to the checkmate, which can involve a comprehensive logical sequence.

Data Processing: We have categorized the data from the validation set into groups based on every 10 moves, with "one move" indicating that each player has made a single move. For instance, a state with 33 moves would fall into the [31, 40) moves category. Given the limited size of the dataset, it is solely utilized as a validation set. We derived our training dataset from a subset of data used in the research by Feng et al. (2023) on ChessGPT. This subset includes games in Portable Game Notation

Table 2: Results on Checkmate in One Move dataset.

Moves	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
Text Acc	19.71	29.03	26.3	28.47	35.15	33.01	36.94	50.0	50.0	33.33
CoI Acc	**41.73**	**39.94**	**32.9**	**40.28**	**40.36**	**53.59**	**68.47**	**83.33**	**75.0**	**66.66**
Img Sim	100	99.93	99.63	99.67	99.93	99.92	99.93	99.9	99.91	99.92
n_train	1815	2200	3392	4428	3616	3043	2930	2618	1752	1134
n_test	137	651	1076	864	384	209	111	42	16	6

(PGN) format from the Pro-player dataset[2] and the CCRL dataset[3]. We meticulously filtered this data to exclude games with non-standard openings, any illegal moves, or those not culminating in checkmate. The final counts of questions in the training and test sets are displayed in the last two rows of Table 2. The final moves, which represent checkmates, are labeled as answers, and the preceding moves as questions. This training set was then converted into SVG format as depicted in Figure 7.

Model Structure: Our SyMLLM is an extension of ChessGPT (Feng et al., 2023), which was fine-tuned on GPT-NeoX-3B (Black et al., 2022) using a vast dataset related to chess games and language. SyMLLM incorporates an image encoder based on the clip-vit-large-patch14 model. The architecture is designed to first generate a symbolic representation in SVG format from the training set, then convert this SVG into an image. This image is subsequently processed by the image encoder, which aids in predicting the checkmate in one move based on the current state of the chessboard.

Training Details: We following Dettmers et al. (2023) applied 4-bit quantization to the LLM's backbone and used a rank-16 LoRA, focusing on training newly added parameters. The fine-tuning process encompassed both the ViT and LLM, utilizing a single A800 GPU, with a batch size of 1 and 8 gradient accumulation steps, spanning over 5 epochs and completed within 8 hours.

Experiment Results: The performance of the SyMLLM on the "Checkmate in One Move" task is summarized in Table 2, showcasing the CoI's high proficiency. The table indicates that SyMLLM achieves an accuracy rate near 100% in generating the correct chessboard state images. This accuracy facilitates the use of CoI reasoning to identify checkmate positions directly from the images, bypassing the need for complex textual reasoning chains. The CoI method significantly outperforms the baseline that relies on text-based inputs. For instance, the accuracy of [11,20) moves has improved to 1.38×, and the accuracy of [61,70) moves has increased to 1.85×.

Results Analysis: When the representation of the chessboard state is consistently accurate, the length of the reasoning chain becomes irrelevant. As the number of moves in a sequence increases, the pool of legal moves narrows, enhancing the likelihood of successfully identifying a checkmate in one move. This is corroborated by the results in the table, where an increase in the number of steps correlates with improved outcomes.

4.3 COMMONSENSE

Task Description: Previous experiments have demonstrated that the Context of Inference (CoI) significantly aids in solving complex logical problems. In this section, we proceed to investigate whether images can enhance the capability of text in common sense reasoning. To this end, we have selected the Location and Unusual tasks for testing. These tasks utilized the same set of 531 event descriptions, but the questions differed. The Location task requires the description of an event, followed by a question about where the event took place. A model's answer is deemed correct only if it includes the provided standard answer. On the other hand, the Unusual task describes a scenario and then asks which parts of the description violate common sense. Given that questions about violations of common sense are open-ended, the accuracy of the answers cannot be directly calcu-

[2] https://www.pgnmentor.com
[3] https://ccrl.chessdom.com

Table 3: Results on the Commonsense Dataset.

	Location Accuracy(%)	Unusual Accuracy(%)
Text	73.63	90.4
CoI with SDXL	77.78	**100**
CoI with DALL·E 3	**82.29**	**100**

lated. Therefore, the study adopted different metrics: if the large-scale model fails to identify the nonsensical elements in the description, the answer is considered incorrect; otherwise, it is correct.

Model Structure: While the SyMLLM is adept at accurately rendering abstract graphics, the generative model creates richly detailed, realistic images. Therefore, we can use SDXL and DALL·E 3 to create common sense images without training a new model. We use LLaVA-13B, a model obtained by finetuning Vicuna-13B and ViT on a text-image dataset, to recognize these common sense images and answer questions. The pure text model we compare against is Vicuna-13B. The CoI method employs the prompt, "Let's solve the problem with the help of an image," to cue the model to use imagery for reasoning.

Experiment Results: As shown in the following table: the use of pictures as a medium can introduce visual common sense information to help improve common sense reasoning problems. For example, using SDXL to generate pictures can improve the accuracy of determining the location of an event by 4.15%. Using the more effective DALL-E 3 to generate pictures improves the accuracy by 9.66%. Using the plain text model, 9.6% of the samples could not find the anti-commonsense points in the textual descriptions, but it is easy to find them after using the pictures.

Results Analysis: As shown in Figure 9 and Figure 10, when solving the same problem, CoI is more intuitive compared to text-only reasoning and is able to supplement the missing details in the text with visual commonsense knowledge to aid the reasoning process. In addition, the strength of the image generation capability directly affects the reasoning efficiency of the CoI; for example, the CoI with DALL-E 3, which generates images of higher quality, is 4.51% more accurate in recognizing the location of an event than the CoI with SDXL.

Question: A woman is waiting to get on the elevator. But the people in the elevator are on fire. Where can this event take place?
Answer: An elevator
Text Only: This event can take place in a dream or a surrealistic setting, as it is a highly unusual and unlikely scenario.
CoI(SDXL): *Let's solve the problem with the help of an image.* This event can take place in a building, specifically in an elevator. The image shows a woman standing in front of an elevator with flames coming out of the elevator's doors. This situation could occur due to a malfunction or an accident, such as an electrical short-circuit or a fire caused by an external source.

CoI(DALL·E 3): *Let's solve the problem with the help of an image.* This event can take place in a building or a facility where elevators are present. *In the image,* the woman is waiting to get on the elevator, but the people inside the elevator are on fire. This situation might occur due to an accident, malfunction, or a supernatural event. The woman's reaction to the situation is crucial, as she needs to take appropriate actions to ensure her safety and the safety of others in the building.

Figure 9: This is an example for the same question, the response of Text only reasoner, Chain of images with original picture and CoI with image generated by stable diffusion. CoI improves the performance of this problem

Question: A group of people are crowded around in a living room talking to one another. A man in the foreground introduces two individuals one appears to be a regular human male the other appears to be an animal. **What is unusual about this description?**

Answer: There is a hairy, furry animal standing upright in a living room with other human beings inside.

Text Only: There is **nothing unusual** about a man introducing an animal to another individual in a living room. This could imply that the situation is not extraordinary or out of the ordinary.

CoI(SDXL): ***Let's solve the problem with the help of an image.*** The unusual aspect of the description is that it describes a man introducing two individuals, one of whom is an animal. This is not a typical scenario in everyday life, as animals are not typically present in social gatherings or introductions. However, in the image, there is a group of people crowded around in a living room, and one man is introducing two individuals, one of whom appears to be an animal.

CoI(DALL·E 3): ***Let's solve the problem with the help of an image.*** The unusual aspect of this description is that the animal is not a typical pet or wildlife but rather a person dressed up as an animal. This creates a unique and unexpected situation in the living room where people are gathered to talk and interact with one another.

Figure 10: This is an example for the same question, the response of Text only reasoner, Chain of images with original picture and CoI with image generated by stable diffusion. CoI improve the performance of this problem

5 Related Work

5.1 Reasoning with LLM

With the emergence of LLMs with surprisingly powerful capabilities, it has become possible to use language models to address various non-traditional natural language tasks, such as some reasoning tasks: solving arithmetic problems, deducting about the commonsense, playing a game (OpenAI, 2023; Srivastava et al., 2022; Qin et al., 2023), etc. This ability is first manifested in the series of GPT model (Radford et al., 2019; Brown et al., 2020), where the models address various tasks without fine-tuning by following the instruction of a paragraph of natural language called "*prompts*". Centered around this prompt-leading generation capability, a series of works have been proposed to enhance the reasoning ability of large models using different forms of prompts: Wei et al. (2022) introduced a chain-of-thought prompt which provides the model with learning examples of the target reasoning task containing the detailed reasoning steps, enabling the model to imitate the examples to generate answers to solve reasoning tasks—this capability is referred to as "*in-context learning*" capability. Further, simply based on tasks described in natural language by the model (rather than examples), it can also stimulate the model's reasoning ability—this is called a 'zero-shot prompt' (Kojima et al., 2022).

At the same time, some non-natural language prompting methods have been proposed. "code prompts" (Gao et al., 2023; Hu et al., 2023) are typical work, using code to help the model achieve better performance on some symbolic reasoning tasks. Liu et al. (2022) further uses the code to operate simulation for some physical situations. Nye et al. (2021) use the intermediate state as the prompt to help the model address arithmetic problems. Chen (2023) use a table as the prompt. The idea behind this kind of work is that, compared to natural language, a more structured prompt can be more helpful for tasks with typical workflows. Unlike these works, we have noticed the limitations of using language and other sequential information to represent problems. We hope to point out that images such as flows, diagrams and illustrations might be another important type of information representation. For some tasks, these images can encode the necessary intermediate or step-by-step information for reasoning with greater information density and in a more appropriate manner.

5.2 Multimodel large language model

The monumental advancements in large language models within the domain of natural language processing have spurred a surge in efforts to adapt these successes to other modalities. Unlike these language models, multimodal models are designed to accept or produce input in various modalities beyond just natural language. These models incorporate additional modules to understand and generate content in formats such as images, voice, or video.

The underlying principle of this work involves integrating the pre-trained language model with another pre-train model for a specific modality, such as images. For instance, Li et al. (2023) proposes to train a lightweight Transformer to align the image representation from an image encoder with the textual representation from an LLM. Drawing inspiration from the instruction tuning techniques used in pure language models, Liu et al. (2023a) employs data generated by GPT-4 to enhance question-answering performance. Alayrac et al. (2022) achieves closer integration of image and text data by incorporating image information into the LLM through cross attention layer instead of embedding. Moreover, Girdhar et al. (2023) proposes a joint embedding across six different modalities, including images, text, audio, depth, terminal, and IMU data, without the need for training on the last four modalities. This is achieved by leveraging the zero-shot capacity of the large image-text models. The above-mentioned models do not possess image-generation capabilities. To make up for this shortcoming, Dong et al. (2023) realize image generation by combining the model proposed by Liu et al. (2023a) with the Stable diffusion (Rombach et al., 2022). Building on this, Wu et al. (2023) further integrates it with the model from Girdhar et al. (2023), utilizing its ability to align multiple modalities to generate and accept multiple modalities. Very recently, GPT-4 (OpenAI, 2023) has demonstrated its astounding ability to understand and generate in the modalities of images and voice. In the examples it showcased, the GPT-4 has already reached a level close to human capabilities, adding both confidence and pressure to the research in this field.

Although our methodology is similar to these works, our motivation leans more towards using images as a tool to assist large models in reasoning (similar to CoT, code, and other descriptions of the intermediate process of reasoning we mentioned in Section 5.1), rather than focusing on its multimodal interaction capabilities. Therefore, compared to the quality of generating text, images, and voice, we are more concerned about whether and in which tasks these inputs and outputs can better assist the model in completing its reasoning.

6 CONCLUSIONS

In this paper, we put forward a Chain-of-Image (CoI) prompting method, which imitates humans' ability to solve complex reasoning problems with the help of images. To accurately generate images to help the CoI reasoning, we propose the SyMLLM framework. We perform experiments on three kinds of datasets: Geometric, Chess and Commonsense, using the SyMLLM framework to demonstrate the power of CoI. We found that CoI uniformly improves the performance over pure-text models without images as an intermediate representation. We believe our CoI method has great potential to help LLMs achieve human-level reasoning ability.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL `https://openreview.net/forum?id=EbMuimAbPbs`.

Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.

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

Wenhu Chen. Large language models are few(1)-shot table reasoners. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1120–1130, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.83. URL `https://aclanthology.org/2023.findings-eacl.83`.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. *arXiv preprint arXiv:2309.11499*, 2023.

Xidong Feng, Yicheng Luo, Ziyan Wang, Hongrui Tang, Mengyue Yang, Kun Shao, David Mguni, Yali Du, and Jun Wang. Chessgpt: Bridging policy learning and language modeling. *arXiv preprint arXiv:2306.09200*, 2023.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models, 2023.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Jack Hessel, Ana Marasović, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. Do androids laugh at electric sheep? Humor "understanding" benchmarks from The New Yorker Caption Contest. In *Proceedings of the ACL*, 2023.

Yi Hu, Haotong Yang, Zhouchen Lin, and Muhan Zhang. Code prompting: a neural symbolic method for complex reasoning in large language models, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

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

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023a.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023b.

Ruibo Liu, Jason Wei, Shixiang Shane Gu, Te-Yen Wu, Soroush Vosoughi, Claire Cui, Denny Zhou, and Andrew M. Dai. Mind's eye: Grounded language model reasoning through simulation, 2022.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models, 2021.

OpenAI. Gpt-4 technical report, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver?, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens, 2023a.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023b.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

Checkmate in One @ BIGBench Example: In the following chess position, find a checkmate-in-one move. 1. f4 Nc6 2. Nf3 Nf6 3. d4 d5 4. e3 Bd7 5. Be2 e6 GPT-4 Response: Yes, generating an image of the board position can help a player to quickly identify threats, patterns, and potential moves.
Chess State Tracking @ BIGBench Example: For each of the following chess games, please complete the notation for the last shown move by filling in the destination square: d2d4 e7e6 e2e3 c7c5 f2f4 b7b6 g1f3 d7d6 f1e2 a7a6 e1g1 g7g6 c2c3 f8g7 c1d2 g8e7 d2e1 e8g8 e3e4 c8b7 b1d2 f7f5 d1b3 b7 GPT-4 Response: Yes, generating an image of the chess board would allow for a more intuitive understanding of the game’s progression and the positions of the pieces.
Geometric Shapes @ BIGBench Example: This SVG path element <path d=”M 31.94,78.63 L 66.91,49.50 L 68.54,41.07 L 61.03,39.02 M 61.03,39.02 L 52.78,44.98 M 52.78,44.98 L 31.94,78.63”/>draws a [sector, heptagon, octagon, circle, rectangle, hexagon, triangle, line, kite, pentagon] GPT-4 Response: Yes, an image would provide a clear, visual representation of the shape, making it easier to identify.
Intersect Geometry @ BIGBench Example: Find the number of intersection points between the shapes and lines specified by the coordinates given. There is a line segment from (-0.1, -2.2) to (-2.4, -2.4). There is a circle centered at (1.8, 1.5) with radius 2.0. How many intersection points are there? GPT-4 Response: Yes, by plotting the line and the circle on a graph, one can easily see where they intersect, especially for those who are more visually inclined.
Logic Grid Puzzle @ BIGBench Example: There are 2 houses next to each other, numbered 1 on the left and 2 on the right. There is one person living in each house. The people in these houses have different characteristics: Each person plays a different sport: one is a soccer player and one is a baseball player. Each person has a different device: one has a radio and one has a gameboy. Clues: 1. The baseball player lives in the first house. 2. The person who has a gameboy does not live in the first house. What is the number of the house where the person who has a gameboy lives? GPT-4 Response: Yes, it can help to visualize the positions of the houses and the characteristics of the people living in them, making it easier to solve the problem.
Logical Deduction @ BIGBench Example: The following paragraphs each describe a set of five objects arranged in a fixed order. The statements are logically consistent within each paragraph. On a shelf, there are five books: a green book, a blue book, a white book, a purple book, and a yellow book. The blue book is to the right of the yellow book. The white book is to the left of the yellow book. The blue book is the second from the right. The purple book is the second from the left. GPT-4 Response: Yes, visualizing the arrangement of the books can make it easier to understand their relative positions and to answer questions about the order.
Matrix Shapes @ BIGBench Example: Compute the kronecker product of a matrix of shape (2,2,3,3) with a matrix of shape (2,2,4,2). Add the result to a matrix of shape (4,4,12,6). Sum the result over the second axis. Compute the kronecker product of the result with a matrix of shape (2,3,3). GPT-4 Response: Yes, generating images can help in visualizing the shapes and dimensions of the matrices involved, making it easier to comprehend and solve the problem.
Movie Dialog Same or Different @ BIGBench Example: The following is a conversation between two people, but the transcript doesn’t mark who said what: ”You’re asking me out.” ”That’s so cute.” ”What’s your name again?” ”Forget it.” —– In the preceding conversation, were the sentences ”That’s so cute.” and ”What’s your name again?” said by the same or different individuals? GPT-4 Response: Yes, it can help to visually organize the conversation and make it easier to track who might be speaking at any given time.
Reasoning About Colored Objects @ BIGBench Example: On the table, you see a bunch of objects arranged in a row: a grey cat toy, a brown paperclip, an orange envelope, a gold notebook, a red teddy bear, and a magenta mug. What is the color of the object directly to the right of the gold object? GPT-4 Response: Yes, it allows for a more intuitive understanding of the spatial arrangement of the objects, making it easier to identify the object to the right of the gold object.
AQuA-RAT @ AGIEval Example: A car is being driven, in a straight line and at a uniform speed, towards the base of a vertical tower. The top of the tower is observed from the car and, in the process, it takes 10 minutes for the angle of elevation to change from 45° to 60°. After how much more time will this car reach the base of the tower? GPT-4 Response: Yes, visual representation can be helpful for this problem. Drawing a diagram with the tower, car’s positions, and angles of elevation can make it easier to apply trigonometric concepts and visualize the distances and relationships between the car and tower.
LogiQA-EN @ AGIEval Example: In the planning of a new district in a township, it was decided to build a special community in the southeast, northwest, centered on the citizen park.These four communities are designated as cultural area, leisure area, commercial area and administrative service area.It is known that the administrative service area is southwest of the cultural area, and the cultural area is southeast of the leisure area. Based on the above statement, which of the following can be derived? GPT-4 Response: Yes, in cases like this where spatial relationships are described, a visual representation can be very helpful. By mapping out the relative positions of the communities, it can provide a clearer understanding of the arrangement and can assist in deriving accurate conclusions based on the provided information.
LSAT-AR @ AGIEval Example: Of the eight students—George, Helen, Irving, Kyle, Lenore, Nina, Olivia, and Robert—in a seminar, exactly six will give individual oral reports during three consecutive days—Monday, Tuesday, and Wednesday. Exactly two reports will be given each day—one in the morning and one in the afternoon—according to the following conditions: Tuesday is the only day on which George can give a report. Neither Olivia nor Robert can give an afternoon report. If Nina gives a report, then on the next day Helen and Irving must both give reports, unless Nina’s report is given on Wednesday.”, ”question”: ”Which one of the following could be the schedule of the students’ reports?” GPT-4 Response: Yes, for this kind of logic-based puzzle, visual representation like a chart or a table can be beneficial. It allows for easier organization of information and constraints, making the process of elimination and arrangement more intuitive.
High School Mathematics @ MMLU Example: If a pentagon P with vertices at (– 2, – 4), (– 4, 1), (–1, 4), (2, 4), and (3, 0) is reflected across the line y = x to get a new pentagon, P’, then one of the vertices of P’ is GPT-4 Response: Yes, by plotting the points and reflecting them across the line, it becomes easier to identify the new position of the vertices and determine which one corresponds to P’.
Location @ NYCC Example: In a tree, a mother bird is tending to her nest and babies. There is another bird in a cage on an upper branch. Where can this event take place? GPT-4 Response: Yes, the key of this problem is that the location should allow for trees where birds can naturally make nests and also accommodate a caged bird. An image could help illustrate this scenario, making it more tangible and perhaps easier to understand, especially for visual learners.
Unusual @ NYCC Example: A group of people are crowded around in a living room talking to one another. A man in the foreground introduces two individuals one appears to be a regular human male the other appears to be an animal. What is unusual about this description? GPT-4 Response: Yes, an image can add context and may help highlight the oddity of an animal being introduced as if it were a person, which is the unusual aspect of this description.

Table 4: Datasets that GPT-4 believes generate images would be helpful.