Cross-Modal Consistency in Large Multimodal Models: An In-Depth Study with GPT-4V

Anonymous ACL submission

Abstract

Recent developments in multimodal methodologies have marked the beginning of an exciting era for models adept at processing diverse data types, encompassing text, audio, and visual content. Models like GPT-4V, which merge computer vision with advanced language processing, exhibit extraordinary proficiency in handling intricate tasks that require a simultaneous understanding of both textual and visual information. Prior research efforts have meticulously evaluated the efficacy of these Vision Large Language Models (VLLMs) in various domains, including object detection, image captioning, and other related fields. However, existing analyses have often suffered from limitations, primarily centering on the isolated evaluation of each modality's performance while neglecting to explore their intricate cross-modal interactions. Specifically, the question of whether these models achieve the same level of accuracy when confronted with identical task instances across different modalities remains unanswered. In this study, we take the initiative to delve into the interaction and comparison among these modalities of interest by introducing a novel concept termed *cross-modal consistency*. Furthermore, we propose a quantitative evaluation framework founded on this concept. Our experimental findings, drawn from a curated collection of parallel vision-language datasets developed by us, unveil a pronounced inconsistency between the vision and language modalities within GPT-4V, despite its portrayal as a unified multimodal model. Our research yields insights into the appropriate utilization of such models and hints at potential avenues for enhancing their design.

1 Introduction

Recent large multimodal models have showcased remarkable capabilities in tasks that require the integration of multiple modalities and sources of information (Huang et al., 2023). Among these, the performance of Vision Large Language Models (VLLMs) (Zhang et al., 2023a; Yang et al., 2023) stands out, thanks to the vast amounts of image and text data available for training and the rapid progress in both computer vision and language modeling. However, due to the distinct training methodologies employed by these models, such as contrastive learning (Radford et al., 2021) and embodied image-language modeling (Driess et al., 2023), and the varying quality of training data for each modality, these networks often exhibit performance disparities across different modalities.

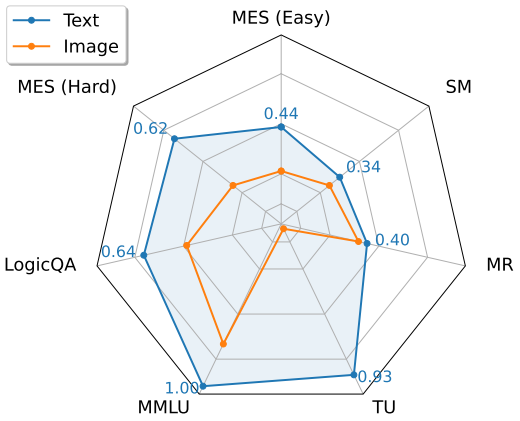

Figure 1: Visualization of the performance gap between the modality of text and image in seven different tasks.

Previous research has extensively evaluated the performance of individual modalities in multimodal systems. For instance, Yang et al. (2023) conducted a thorough assessment of GPT-4V's vision understanding capabilities, and Chen et al. (2023) analyzed model's decision-making abilities. However, assessing a model's performance on each individual modality in isolation does not fully evaluate its true multimodal abilities. It is possible, for example, for a model to excel in numerous vision

tasks but still lag significantly behind in language understanding. Moreover, simply testing performance on individual tasks provides no insight into whether and how each modality of the model influences the others. Unfortunately, the cross-modality relationship is frequently overlooked in the aforementioned research.

In this study, we go beyond the traditional approach of simply evaluating multimodal systems through separate downstream tasks and reporting their scores. Our focus is primarily on measuring the inherent *differences* in capabilities between various modalities, with special attention to vision and language, given their prominence among other modalities. To enable a comprehensive analysis, we introduce the concept of *cross-modal consistency*, complete with a formal definition and an evaluation framework. We consider cross-modal consistency to be an essential element in the design of complex multimodal systems with neural components, as it guarantees coherence and reliability in the system's performance. This is crucial for both interpretability and for fostering user trust.

We subsequently construct a comprehensive vision-language parallel dataset encompassing seven tasks, each designed to highlight different facets of vision and language capabilities. This dataset serves as a tool for evaluating the vision-language consistency of VLLMs. Our experiments with the GPT-4V model on the dataset reveal significant inconsistencies between its vision and language capabilities. The results indicate that its performance varies considerably depending on whether the same task instance is prompted in one modality versus the other.

Our contributions are: (1) We introduce the novel concept of cross-modal consistency, along with a comprehensive evaluation framework. This approach transcends traditional assessment methods for multimodal models, which typically evaluate each modality in isolation. (2) We develop and release seven diverse datasets, carefully designed for vision-language consistency evaluation, opening up opportunities to exploit these datasets in future research. (3) Our experiments on GPT-4V reveal a significant disparity between vision and language abilities within such a system, prompting the introduction of the Vision-Depicting-Prompting (VDP) method as a potential remedy. Our findings offer valuable guidance for more effective future use of such multimodal models.

2 Related Work

A substantial amount of effort has been dedicated to meticulous evaluation of large multimodal models such as GPT-4V. To assess the capabilities of these models across all their modalities, a wide array of tasks has been tested. E.g., researchers have scrutinized GPT-4V's aptitude in solving problems within specialized domains, including biomedicine (Liu et al., 2023b), medical applications (Wu et al., 2023), and autonomous driving (Wen et al., 2023), employing intricate image inputs. Beyond these domain-specific evaluations, more general skills like chart image understanding (Liu et al., 2023a) and optical character recognition (Shi et al., 2023) have also been analyzed. However, these evaluations often focus solely on performance metrics for each test dataset, with little or no exploration of the relative capability gaps between vision and language. In this study, our primary emphasis lies in uncovering the relative *disparities* in the abilities of multimodal models across their various modalities, rather than merely assessing absolute performance within specific tasks.

Despite the lack of cross-modal analysis for multimodal models, previous research has delved into examining cross-lingual abilities in Large Language Models (LLMs). For example, by translating task instances into different languages and analyzing the pairwise results, Zhang et al. (2023b) demonstrated that models like GPT-3.5, primarily trained on English text corpora, exhibit disparities in their performance across various tasks when prompted with different languages. Specifically, these LLMs display a bias toward English. Taking inspiration from such studies, we extend our research to encompass consistency analysis across various modalities, recognizing that different languages can be regarded as distinct modalities as well. Our generalized framework sheds light on the underlying principles governing the **consistency** of multimodal models when confronted with tasks in diverse modalities, thereby contributing to a deeper understanding of their capabilities and limitations.

3 Preliminaries and Key Concepts

As "consistency" can carry different interpretations within the specific context we are addressing, a formal definition of the concept of cross-modal consistency for multimodal models is warranted. To that end, we establish an instance of task t, represented as the paired value (d_a, q). Here, d_a

represents a data element from the input space $\mathcal{D}_a$ corresponding to modality a, while $q \in \mathcal{Q}$ represents the abstract query, often presented in the form of a question pertinent to the task at hand. A task set within modality a is then constituted by combining certain data elements from modality a with the queries q, which can be denoted as $S_{t,a} = \{(d_a^{(1)}, q), (d_a^{(2)}, q), (d_a^{(3)}, q), \ldots\}$. When the queries q are held constant, and elements $d_b \in \mathcal{D}_b$ in another modality b are gathered, we obtain the corresponding task set in another modality, denoted as $S_{q,b}$. In essence, the task t embodies the task-specific queries, encompassing, e.g., activities such as solving equations, translation, question answering, etc. Meanwhile, the data elements d_m may take the form of equation instances or question descriptions within modality m, which can involve the modalities of image, text, or speech.

We introduce the concept of a 'converter,' a function $K_{a,b} : \mathcal{D}_a \mapsto \mathcal{D}_b$ which maps data elements from modality a to b. While there exist various methods for converting data between modalities (e.g., from language to vision through taking a picture), we are specifically interested in converters that preserve information necessary for solving a given task with query q, denoted as $K^q_{a,b}$. Information-preserving converters are distinctive, as the correct answer for a given task instance (d, q) depends solely on the information within d rather than its modality. Therefore, both (d_a, q) and $(K^q_{a,b}(d_a), q)$ are guaranteed to share the same gold label. In this work, we assume the existence of K^q for every $q \in \mathcal{Q}$, but finding such a converter is beyond the scope of this paper. Inter-modality conversion may be challenging for certain modalities. Some tasks may involve aspects of information, such as emotions in speech or nuanced visual perception in images, that cannot be easily preserved during conversion. We design our experiments with tasks where a K^q clearly exists.

A multimodal model can be conceptualized as a function, denoted $M : \mathcal{D} \times \mathcal{Q} \mapsto \mathcal{Y}$, mapping data elements and queries to an answer. Here, $\mathcal{D}$ represents the collective space encompassing all the modalities of interest, formally $\mathcal{D} = \bigcup_m D_m$, where m spans over all relevant modalities. On the other hand, the answer space $\mathcal{Y}$ refers to a unified and structured representation, which, in the case of GPT-4V, assumes the form of text.

A model M is said to exhibit consistency between modalities a and b provided:

$$M(d_a, q) = M(K^q_{a,b}(d_a), q), \forall d_a \in D_a,\ q \in \mathcal{Q}$$

In other words, M is *consistent* if its output is invariant under any modality transformation K^q which preserves all essential information necessary for solving the task associated with query q. E.g., consider solving mathematical equations. A model which solves this task is consistent across the text and image modalities if neither transcribing the equation from image to text, nor imaging an equation presented as text, changes the model's output.

In short, a consistent model should remain agnostic to the modality of the task instance and yield identical results as long as an equivalent amount of information is provided, reflecting its capacity to handle multimodal data seamlessly.

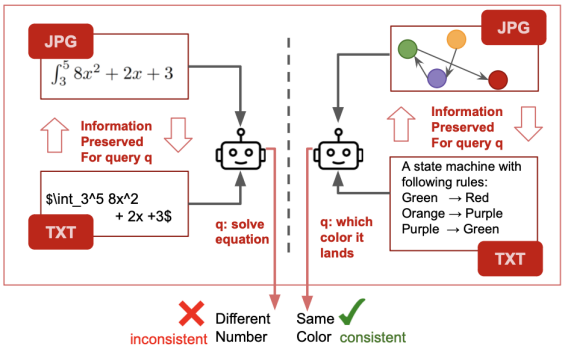

Figure 2: Illustration of the concept of cross-modal consistency. A consistent model (right) applies the same internal reasoning to task instances with identical information, regardless of the encoding modality, leading to consistent outcomes. In contrast, an inconsistent model displays significant behavioral changes in response to different input modalities, resulting in varying outcomes as the modality alters.

4 Method

In this section, we describe our method for testing cross-modal consistency. We establish a quantitative evaluation framework, with a focus on the vision-language cross-modality. We provide a description of our methodology and the specific metrics we propose for evaluation.

4.1 WorkFlow

For an instance set of a given task t in modality a, denoted as $S_{t,a} = \{(d_a^{(1)}, q), (d_a^{(2)}, q), (d_a^{(3)}, q), \cdots\}$, our first step involves constructing a parallel instance set $S_{t,b}$ in modality b using an information-preserving converter $K^q_{a,b}$. We do so by apply-

ing $K^q_{a,b}$ to each data object $d^{(i)}_a$ to get the object $d^{(i)}_b := K^q_{a,b}(d_a)$ in modality b. By doing so, each paired instance $(d^{(i)}_a, q)$ and $(d^{(i)}_b, q)$ shares the same gold label since the information in d is preserved for the task with query q. In the context of analyzing the vision and language modalities, our converter is comprised of an optical character recognition (OCR) system combined with human verification for converting images to text, and screenshot software for converting text into images. We carefully select tasks where the information required for solving the task can be fully retained through the utilization of this converter, as exemplified by mathematical equation solving.

Next, we independently apply the model M to each pair of instances $(d^{(i)}_a, q)$ and $(d^{(i)}_b, q)$ to obtain pairwise results $M(d^{(i)}_a, q)$ and $M(d^{(i)}_b, q)$.

4.2 Metrics

We introduce our task consistency score C_t based on these pairwise instances:

$$C_t = \frac{1}{n} \sum_{i=1}^{n} c^i_M \tag{1}$$

where

$$c^i_M = \begin{cases} 1, & \text{if } M(d^{(i)}_a, q) = M(d^{(i)}_b, q) \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

In essence, C_t is the proportion of instances for which model M has consistent performance on the given task, between modalities a and b.

5 Experiments

5.1 Data Construction

Since there is currently no existing parallel vision-language task dataset, we create our own datasets for both our experiments and also to facilitate future research endeavors. Following the approach outlined in Section 4.1, we meticulously selected seven tasks that gauge various facets of Vision-Large Language models. For each of these tasks, we ensure that data instances can be transformed between image and text formats while preserving all task-related information, utilizing a straightforward converter (e.g., OCR). Recognizing that a flawless converter does not exist in practice, we undertake the manual verification of each converted data instance to prevent any potential errors during the conversion process. We will make our dataset available for use by the research community in the final version of our paper.

5.1.1 Task Description.

Math Equation Solving. Mathematical reasoning stands as a cornerstone of multi-modal models' capabilities. Mathematical problems typically involve equations presented in a visual format, offering a clear depiction of intricate symbols and notations. Given that formulas can be seamlessly converted to text formats like LaTeX without losing any essential information for solving these equations, constructing a parallel dataset for such tasks is a natural fit for analyzing cross-modal consistency. For our dataset, we source math questions with equations from two distinct origins, each representing varying levels of difficulty. For low difficulty levels, we extract 901 high school-level mathematical questions in LaTeX (text) format from MATH dataset (Hendrycks et al., 2021b), rendering each question using a LaTeX compiler to generate corresponding image data. To introduce a greater level of complexity, we gathered 50 college-level calculus questions, along with their corresponding answers, using the same procedure. Consequently, we paired all the image-based math questions with their corresponding text representations to create our comprehensive equation-solving dataset, encompassing both easy and challenging questions. An illustrative example of this dataset can be found in Figure 3, and detailed data samples are available in Appendix A and Appendix B .

Logical Reasoning. To assess the vision-language consistency in logical reasoning abilities for the VLLMs, we employ two distinct datasets: GSM8K (Cobbe et al., 2021) and LogicQA (Liu et al., 2020). GSM8K comprises 8,500 question instances in text format, with each instance representing a problem description in English text paired with a labeled answer. We transform the text into images by capturing screenshots of the rendered text with an appropriate font size and layout. Similarly, LogicQA consists of 8,678 more challenging questions presented in text format and each is converted into an image by us. Subsequently, we pair these resulting images with the original text files, creating a parallel dataset that enables the exploration of this task in both image and text modalities. An illustrative example of this dataset construction can be found in Figure 3, and detailed data samples are available in Appendix F and Appendix C.

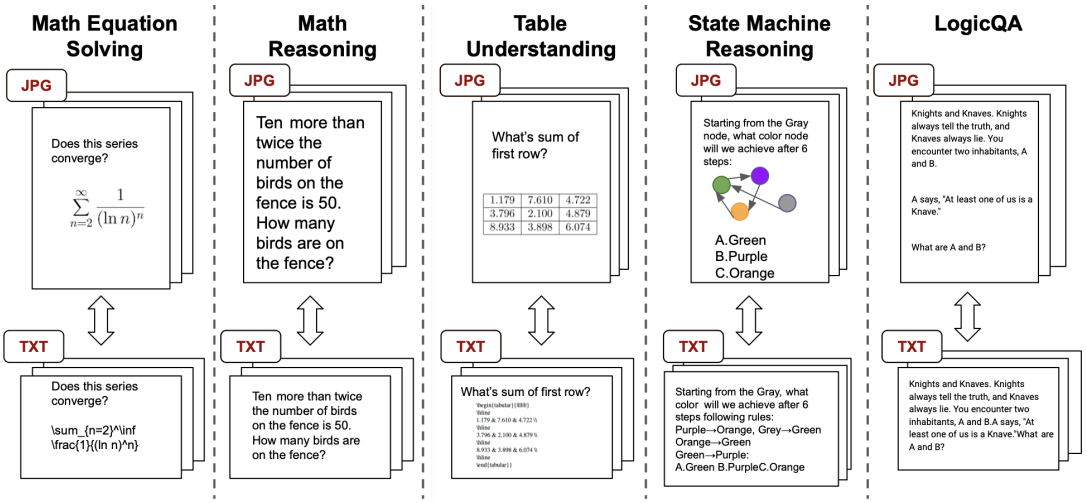

Figure 3: An Overview of the Components of Our Vision-Language Consistency Dataset. Data instances are presented in pairs, featuring one in the vision modality and another in the text modality. Notably, Math Equation Solving dataset encompasses two segments, each representing different difficulty levels.

Table Understanding. Tables, commonly encountered in everyday life, are often presented as images, and the effective extraction of information from them is vital for various tasks. As well-structured table images can be easily converted into LaTeX text, they serve as an excellent choice for conducting vision-language consistency analysis. To facilitate this analysis, we creat 30 distinct tables in LaTeX, each featuring multiple rows and columns, with numerical values in each cell. Our task revolves around accurately summing the numbers within a given row and column. We provide parallel task instances in both LaTeX text and rendered images, as illustrated in Appendix E

State Machine Reasoning. State machines, which can be effectively visualized as graphs or represented through text with transition rules, serve as an ideal test bed for vision-language consistency in simple computational capabilities of VLLMs. Our approach involves generating images of state machines with varying total numbers of nodes (states). Each node in the state machine is assigned a distinct color and features precisely one outgoing edge, ensuring a unique path and solution. The questions we formulate are of the form, "Starting from the color grey, after n steps, which color will we end up in?" Here, n is a variable that we select. Additionally, we generate a text version of these state machines by listing out all the transition rules corresponding to the arrows. To prevent any form of cheating by looking at the last state in the text, we shuffle the order of the rules. We create state machines with different numbers of states and questions with varying numbers of steps, to introduce varying difficulty levels. The data samples can be seen in Appendix G.

Reading Comprehension. To assess the model's consistency in comprehending lengthy English paragraphs across vision and language modalities, we provide the model with the same text content in two different formats: plain text and images of the text. We employ the test part of the Massive Multitask Language Understanding (Hendrycks et al., 2021a), or MMLU dataset as our source, which includes 1,477 extensive text passages, each accompanied by multiple-choice questions designed to evaluate the comprehension of the text content. For this dataset, we convert each text instance into an image by rendering the text into a PDF before converting it to a JPG image. Detailed data samples can be found in the Appendix D.

5.2 Experiment Details.

We apply our framework and constructed datasets to evaluate the cross-modal consistency of the OpenAI GPT-4V model, known for its proficiency in both vision and language modalities. Given the limited daily access to prompt this model, our experiments were conducted on a randomly selected subset of 50 samples from each dataset. We select the GPT-4V classical mode, which does not include additional plug-ins and employs a relatively low decoding temperature to minimize variance in its output. To ensure a fair comparison of capa-

Task	Modal	Acc	Consistency
MES(Easy)	Text	0.44	0.72
	Image	0.24 ⇓	
MES (Hard)	Text	0.62	0.62
	Image	0.28 ⇓	
LogicQA	Text	0.64	0.64
	Image	0.44 ⇓	
MMLU	Text	1.00	0.74
	Image	0.74 ⇓	
TU	Text	0.93	0.10
	Image	0.03 ⇓	
MR	Text	0.40	0.92
	Image	0.36	
State Machine	Text	0.34	0.67
	Image	0.28	

Table 1: Test results for vision-language consistency datasets. MES stands for Math Equation Solving, TU stands for Table Understanding and MR stands for math reasoning. The symbol ⇓ denotes a sizeable decrease in accuracy (greater than 10%) when input is in the image format.

bilities between the two modalities, we embedded the query questions into the image and exclusively used images for prompting. This avoids the involvement of any text input when testing the vision modality. Additionally, to prevent the model from performing reasoning steps in text and introducing unintended modality conversions, we explicitly instructed the model to output answers without any reasoning steps. Our results are manually collected for pairwise data instances, and we calculate the consistency scores based on the methodology outlined in Section 4.1.

5.3 Main Results

The main outcomes of our assessments across seven distinct datasets are outlined in Table 1. Notably, even though the input contains an equivalent amount of information necessary for task completion, substantial disparities emerge between image and text input formats. This phenomenon occurs even in tasks where images are conventionally considered to offer a more vivid and intuitive representation from a human perspective.

We note that consistency, being based on response agreement between modalities, can be high or low regardless of per modality accuracy. The highest consistency (0.92) is observed for math reasoning even though both modalities have a relatively low accuracy (≤ 0.40). By contrast, the consistency drops to 0.64 on logical reasoning (LogicQA) on which the individual modalities have higher accuracy (≥ 0.44).

For tasks that involve intricate reasoning steps, including equation solving, math/logical reasoning, and state machine reasoning, we observe relatively low accuracy even when the input is presented in pure text format. These tasks align with areas where the model generally struggles. When the input modality shifts to using images, the proficiency in solving such tasks deteriorates further, resulting in a noticeable drop in performance, despite the fact that the images contain an equal amount of information. This emphasizes the substantial inconsistency in task-solving across modalities and highlights the model's superior ability in one modality (Language) compared to the other (Vision).

On the other hand, for tasks primarily focused on extracting information from provided content and comprehending that information, such as Language Understanding and Table Understanding, we witness near-perfect performance when the model is prompted with text input. However, a more significant drop in accuracy (up to 90%) is observed in such tasks when the input modality shifts to images. This indicates that the change in modality significantly impacts the model's processing capabilities, providing strong evidence of the inconsistency of the model.

In conclusion, in multimodal systems like GPT-4V, the language modality demonstrates a **dominant** advantage over vision modality, when tasks are tackled in text format, despite the presence of the same information in image format. This strongly suggests a *non-consistent* cross-modal behavior within the network. While each modality exhibits varying levels of task-solving and reasoning capabilities, the inconsistency across modalities is observed across tasks regardless of the accuracy level of each modality for the task in hand.

5.4 Ablation Study on Content Extraction from Images

As solving tasks in image format inevitably requires accessing essential information from the images, we conducted additional experiments to investigate whether the performance gap is attributable to the model's inability to access information. To address this, we conducted one-step Optical Character Recognition (OCR) using the model's own network on all instances of tasks that exhibited a significant performance gap between image and text. Specifically, for each image input (indicated

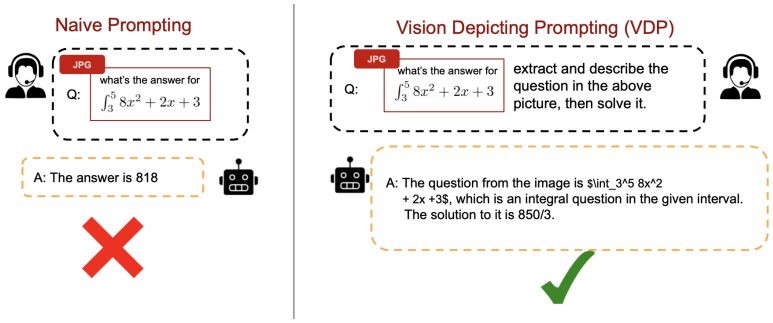

Figure 4: Overview of the VDP Method: The left part illustrates the conventional approach to prompting vision tasks, while the right part demonstrates VDP in comparison.

by the red arrow in Table 1), we prompt the model with the instruction 'extract the exact content in the image' and compare the results with the original input to determine if they match. This approach allows us to eliminate the possibility that the performance issues in image format are due to the model's inability to correctly recognize the input.

As shown in Table 2, OCR accuracy approaches nearly 100% for all instances of LogicQA, MMLU, and Table Understanding tasks. This suggests that the model faces no difficulties in accurately extracting information, such as numbers from each row and column in table images. The substantial gap (up to 90%) in accuracy (Table 1) between images and text can be attributed solely to the model's internal reasoning processes for each modality. This underscores the inconsistent internal reasoning employed by the model when presented with the same content in different modalities.

In contrast, we observe lower OCR accuracy for Math equation-solving inputs, as complex math equations pose challenges for accurate recognition and extraction. To isolate and distinguish the source of inconsistency – inaccurate recognition of image data *or* poor actual internal reasoning, we report *conditional consistency scores* for image instances given correct versus incorrect OCR results. From Table 3, it becomes evident that there is no direct correlation between consistency scores and direct OCR accuracy. This further bolsters our claim that such models simply exhibit distinct (and inconsistent!) internal behaviors under different modalities.

6 Vision-Depicting-Prompting (VDP)

As shown in Section 5.3, for the same task, VLLMs such as GPT-4V perform much better when questions are presented in text format, even when the information can be completely extracted from the image instances. Inspired by these findings, we propose a novel method of *Vision-depicting-prompting* (VDP) for improving model's reasoning ability through image context. We now explain VDP.

DataSet	OCR Accuracy
MES (Easy)	0.68
MES (Hard)	0.76
LogicQA	0.98
MMLU	0.98
TU	1.00

Table 2: Result of performing OCR on the all images of experimented task instances.

DataSet	YConsistency	NConsistency
MES (Easy)	0.70	0.75
MES (Hard)	0.66	0.58

Table 3: Conditional vision-language consistency score given the OCR results. The term 'YConsistency' refers to the consistency given OCR outputs are correct. Conversely, 'NConsistency' denotes the consistency score given incorrect OCR outputs.

6.1 Prompting Details

In the case of a task instance presented in image format, VDP diverges from directly soliciting an answer solely based on the image input, as illustrated in Figure 4. Instead, we adopt a two-step process: we first prompt the model to extract and articulate the description of the image task using textual language. This aims to maximize the transformation of the image signal into a text signal, recognizing the inherently stronger reasoning abilities associated with text information, as demonstrated earlier. Subsequently, we prompt the model to provide an answer, taking into account both the text description of the task and the original image input, as depicted in Figure 4.

Task	Modality	Prompt	Acc	Consistency
MES (Easy)	text	naive	0.44	—
	image	naive	0.24	0.72
		VDP	**0.48** ⇑	0.72
MES (Hard)	text	naive	0.62	—
	image	naive	0.28	0.62
		VDP	**0.50** ⇑	**0.76** ⇑
LogicQA	text	naive	0.64	—
	image	naive	0.44	0.64
		VDP	**0.56** ⇑	**0.80** ⇑
MMLU	text	naive	1.00	—
	image	naive	0.74	0.74
		VDP	**0.98** ⇑	**0.98** ⇑
TU	text	naive	0.93	—
	image	naive	0.03	0.10
		VDP	**0.93** ⇑	**0.90** ⇑

Table 4: Result of VDP prompting. MES stands for Math Equation Solving and TU stands for Table Understanding. ⇑ represents an improvement of more than 10% using VDP.

Unlike previous research that sought to enhance the reasoning abilities of multimodal models by augmenting input images with supplementary text (Lin et al., 2022; Hu et al., 2023), VDP does not focus on information augmentation. Particularly in the task instances designed for our study, images already contain all the necessary information required to complete the task. Therefore, converting these images into text format does not provide any additional information that aids in solving the task. Instead, VDP is rooted in the observation that textual signals can significantly stimulate a model's reasoning capability as model has a bias towards language modality. Instead, VDP is based on the observation that textual signals can significantly stimulate a model's reasoning capability, given the model's inherent bias toward the language modality. VDP achieves this by explicitly extracting textual information from the images, thus directly leveraging the model's language processing capabilities more effectively.

6.2 Experiment Results for VDP

We apply VDP to five of the tasks previously examined in Section 5, where these tasks demonstrate notable performance disparities between image and text inputs. We therefore investigate whether VDP can effectively bridge the performance gap between modalities on such tasks. The outcomes are detailed in Table 4.

Remarkably, we observe a substantial improvement in accuracy exceeding 12% when solving problems within the realm of vision modalities using VDP, as compared to naive prompting. In tasks requiring reasoning abilities, we note an average accuracy enhancement of 19%. However, the overall performance still lags behind that of text-based prompting. This discrepancy can likely be attributed to the challenges in accurately depicting and extracting information from objects within images during VDP. In contrast, an impressive average increase of 57% in accuracy is observed in tasks centered around understanding (TU and MMLU). Particularly, in the case of table understanding, we witness a remarkable 90% boost in accuracy, particularly when the table's content is extracted before any necessary calculations are applied. For these tasks, we find that performance eventually reaches parity with text-based prompting, underscoring the effectiveness of VDP, particularly in tasks that involve a deeper understanding of the information within the input instances.

Furthermore, there is a substantial increase in the consistency score with VDP compared to prompting with plain images (naive prompting), e.g., from 0.64 to 0.80 on LogicQA and from 0.10 to 0.90 on TU. These results reinforce our hypothesis that models such as GPT-4V exhibit varied and often inconsistent reasoning capabilities across different modalities and underscore the effectiveness of our VDP approach for enhancing consistency. Properly addressing such disparities between modalities as done by our VDP approach can also help to improve the performance in solving the tasks.

7 Conclusion

In this study, we performed a systematic analysis of the consistency across modalities in multimodal systems. Our results demonstrate that models such as GPT-4V maintain a relatively independent internal representation of reasoning between visual and textual signals, as evidenced by results we obtained on our datasets which we specially designed for the tasks. Notably, GPT-4V exhibits superior performance in language modeling compared to reasoning within a visual context. These findings offer valuable insights into the potential applications of such multimodal systems and highlight the need for more integrated system designs. Furthermore, we introduce a Vision-depicting-Prompting solution to effectively address this inconsistency.

Limitations

While our method is straightforward and effective in revealing inconsistency across modalities, it does encounter challenges when applied to certain existing tasks. Obtaining an information-preserving converter from one modality to another can prove difficult for specific tasks, such as detecting emotions from speech. Consequently, we may not always be able to readily convert the modality of every given dataset and evaluate the cross-modal consistency of these tasks. However, it is important to note that this limitation should not undermine the value of our approach. Our method provides a general framework for assessing cross-modal behavior, and there exist numerous tasks that can be easily converted across modalities without any loss of information, as demonstrated in our constructed datasets. By testing on such tasks, we can gain a comprehensive understanding of a model's cross-modal behavior.

Ethical Consideration

Our exploration of modality consistency serves as a valuable means to enhance the transparency of multimodal models and gain a profound comprehension of their behavior. By delving into the alignment of model responses across diverse modalities, we uncover intricate insights into the decision-making processes and rationale behind their actions. This comprehensive understanding not only instills confidence in the outcomes generated by these models but also significantly enhances their overall interpretability. Transparency in this context becomes essential not only for establishing trust when these models are integral to pivotal decision-making processes but also for addressing ethical and societal implications. As we unravel the intricacies of multimodal reasoning, it underscores the necessity for continuous ethical contemplation and the implementation of proactive measures to address potential challenges arising from advanced multimodal models.

References

Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. 2023. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. Palm-e: An embodied multimodal language model.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the math dataset.

Yushi Hu, Hang Hua, Zhengyuan Yang, Weijia Shi, Noah A Smith, and Jiebo Luo. 2023. Promptcap: Prompt-guided task-aware image captioning.

Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. 2023. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. Revive: Regional visual representation matters in knowledge-based visual question answering.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. 2023a. Mmc: Advancing multimodal chart understanding with large-scale instruction tuning.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning.

Zhengliang Liu, Hanqi Jiang, Tianyang Zhong, Zihao Wu, Chong Ma, Yiwei Li, Xiaowei Yu, Yutong Zhang, Yi Pan, Peng Shu, Yanjun Lyu, Lu Zhang, Junjie Yao, Peixin Dong, Chao Cao, Zhenxiang Xiao, Jiaqi Wang, Huan Zhao, Shaochen Xu, Yaonai Wei, Jingyuan Chen, Haixing Dai, Peilong Wang, Hao He, Zewei Wang, Xinyu Wang, Xu Zhang, Lin Zhao, Yiheng Liu, Kai Zhang, Liheng Yan, Lichao Sun, Jun Liu, Ning Qiang, Bao Ge, Xiaoyan Cai, Shijie Zhao, Xintao Hu, Yixuan Yuan, Gang Li, Shu Zhang, Xin Zhang, Xi Jiang, Tuo Zhang, Dinggang Shen, Quanzheng Li, Wei Liu, Xiang Li, Dajiang Zhu, and Tianming Liu. 2023b. Holistic evaluation of gpt-4v for biomedical imaging.

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. 2021. Learning transferable visual models from natural language supervision.

Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. 2023. Exploring ocr capabilities of gpt-4v(ision) : A quantitative and in-depth evaluation.

Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, Zheng Zhu, Shaoyan Sun, Yeqi Bai, Xinyu Cai, Min Dou, Shuanglu Hu, Botian Shi, and Yu Qiao. 2023. On the road with gpt-4v(ision): Early explorations of visual-language model on autonomous driving.

Chaoyi Wu, Jiayu Lei, Qiaoyu Zheng, Weike Zhao, Weixiong Lin, Xiaoman Zhang, Xiao Zhou, Ziheng Zhao, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Can gpt-4v(ision) serve medical applications? case studies on gpt-4v for multimodal medical diagnosis.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. 2023. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2023a. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023b. Don't trust gpt when your question is not in english. *arXiv preprint arXiv:2305.16339*.

A Math Equation Solving (Easy) Dataset

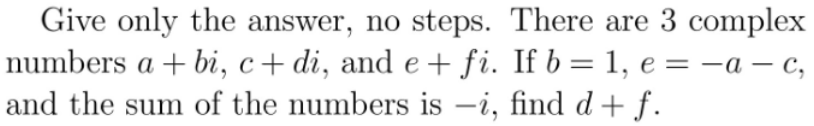

Figure 5: Sample 1 of Math Equation Solving (Easy) Dataset: Image.

Text:	Give only the answer, no steps. Find the largest value of c such that 1 is in the range of $f(x)=x ^ 2-5x+c$.

Table 5: Sample 1 of Math Equation Solving (Easy) Dataset: Text

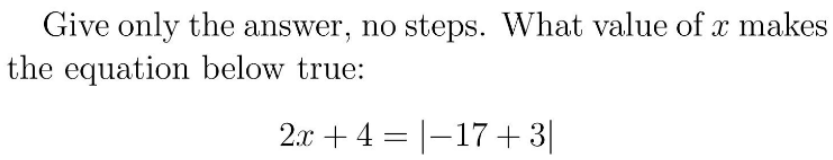

Figure 6: Sample 2 of Math Equation Solving (Easy) Dataset: Image.

Text:	Give only the answer, no steps. What value of x makes the equation below true: $$2x + 4 = \|-17 + 3\|$$

Table 6: Sample 2 of Math Equation Solving (Easy) Dataset: Text

B Math Equation Solving (Hard) Dataset

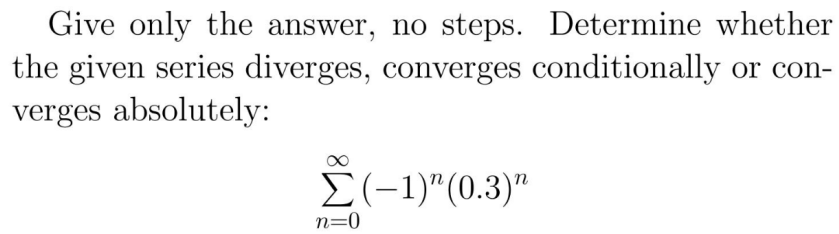

Figure 7: Sample 1 of Math Equation Solving (Hard) Dataset: Image.

Text:	Give only the answer, no steps. Determine whether the given series diverges, converges conditionally or converges absolutely: $$ \sum_{n=0}^ {\infty}(-1)^ n(0.3)^ n$$

Table 7: Sample 1 of Math Equation Solving (Hard) Dataset: Text

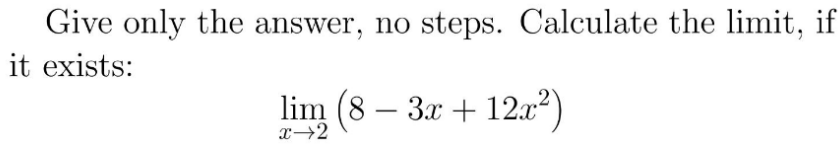

Figure 8: Sample 2 of Math Equation Solving (Hard) Dataset: Image.

Text:	Give only the answer, no steps. Calculate the limit, if it exists: $$ \lim_{ x \rightarrow 2 } \left (8-3 x+12 xˆ 2 \right)$$

Table 8: Sample 2 of Math Equation Solving (Hard) Dataset: Text

C LogicQA Dataset

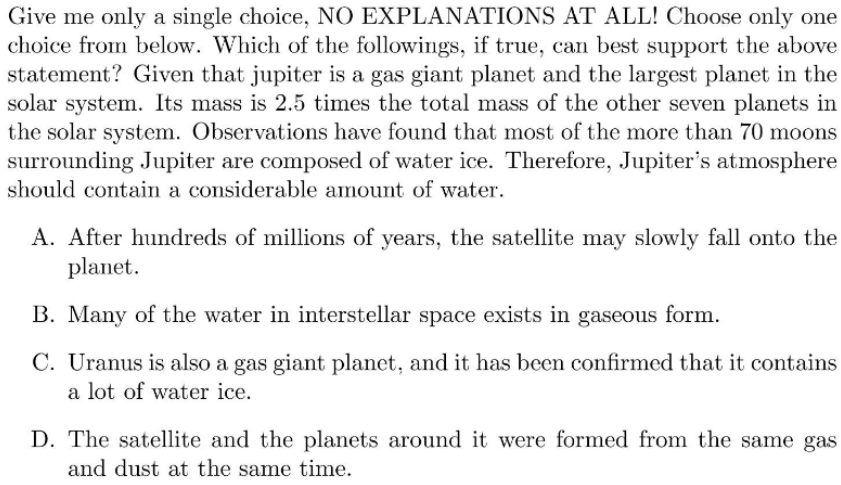

Figure 9: Sample 1 of LogicQA Dataset: Image.

Text:	Give me only a single choice, NO EXPLANATIONS AT ALL! Choose only one choice from below. Which of the followings, if true, can best support the above statement? Given that jupiter is a gas giant planet and the largest planet in the solar system. Its mass is 2.5 times the total mass of the other seven planets in the solar system. Observations have found that most of the more than 70 moons surrounding Jupiter are composed of water ice. Therefore, Jupiter's atmosphere should contain a considerable amount of water. A. After hundreds of millions of years, the satellite may slowly fall onto the planet. B. Many of the water in interstellar space exists in gaseous form. C. Uranus is also a gas giant planet, and it has been confirmed that it contains a lot of water ice. D. The satellite and the planets around it were formed from the same gas and dust at the same time.

Table 9: Sample 1 of LogicQA Dataset: Text

Give me only a single choice, NO EXPLANATIONS AT ALL! Choose only one choice from below. Which of the followings can be infered Given that all Anxi people are vegetarians, while all Zhenyuan people are ascetics. Ascetics and vegetarians are like fire and water, and there is no conflict. Guo Shu is an ascetic.

A. Guo Shu is from Zhenyuan

B. Guo Shu is not from Zhenyuan

C. Guo Shu is from Anxi

D. Guo Shu is not from Anxi

Figure 10: Sample 2 of LogicQA Dataset: Image.

Text:	Give me only a single choice, NO EXPLANATIONS AT ALL! Choose only one choice from below. Which of the followings can be infered Given that all Anxi people are vegetarians, while all Zhenyuan people are ascetics. Ascetics and vegetarians are like fire and water, and there is no conflict. Guo Shu is an ascetic. A. Guo Shu is from Zhenyuan B. Guo Shu is not from Zhenyuan C. Guo Shu is from Anxi D. Guo Shu is not from Anxi

Table 10: Sample 2 of LogicQA Dataset: Text

D MMLU Dataset

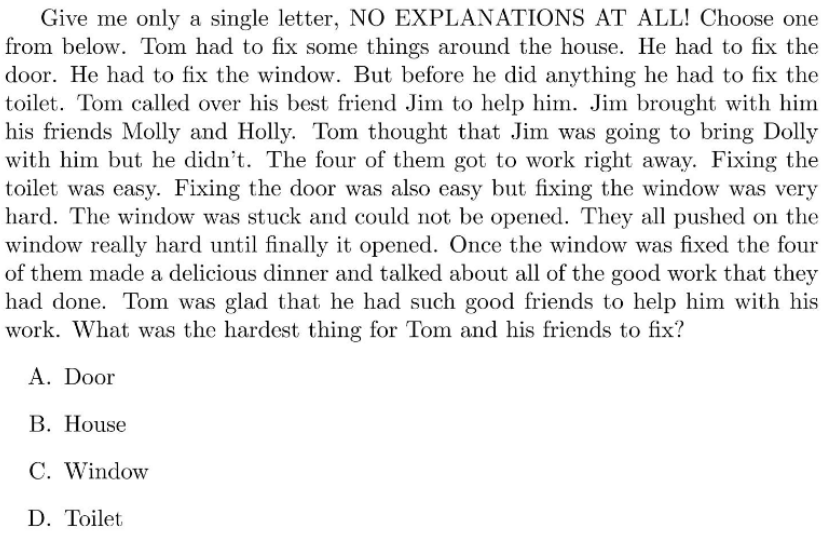

Figure 11: Sample 1 of MMLU Dataset: Image.

Text:	Give me only a single letter, NO EXPLANATIONS AT ALL! Choose one from below. Tom had to fix some things around the house. He had to fix the door. He had to fix the window. But before he did anything he had to fix the toilet. Tom called over his best friend Jim to help him. Jim brought with him his friends Molly and Holly. Tom thought that Jim was going to bring Dolly with him but he didn't. The four of them got to work right away. Fixing the toilet was easy. Fixing the door was also easy but fixing the window was very hard. The window was stuck and could not be opened. They all pushed on the window really hard until finally it opened. Once the window was fixed the four of them made a delicious dinner and talked about all of the good work that they had done. Tom was glad that he had such good friends to help him with his work. What was the hardest thing for Tom and his friends to fix? A. Door B. House C. Window D. Toilet

Table 11: Sample 1 of MMLU Dataset: Text

Give me only a single letter, NO EXPLANATIONS AT ALL! Choose one from below. Lisa has a pet cat named Whiskers. Whiskers is black with a white spot on her chest. Whiskers also has white paws that look like little white mittens. Whiskers likes to sleep in the sun on her favorite chair. Whiskers also likes to drink creamy milk. Lisa is excited because on Saturday, Whiskers turns two years old. After school on Friday, Lisa rushes to the pet store. She wants to buy Whiskers' birthday presents. Last year, she gave Whiskers a play mouse and a blue feather. For this birthday, Lisa is going to give Whiskers a red ball of yarn and a bowl with a picture of a cat on the side. The picture is of a black cat. It looks a lot like Whiskers. What does Whiskers like to do?

A. Sleep in the sun and drink creamy milk

B. Play

C. Drink

D. Sleep

Figure 12: Sample 2 of MMLU Dataset: Image.

Text:	Give me only a single letter, NO EXPLANATIONS AT ALL! Choose one from below. Lisa has a pet cat named Whiskers. Whiskers is black with a white spot on her chest. Whiskers also has white paws that look like little white mittens. Whiskers likes to sleep in the sun on her favorite chair. Whiskers also likes to drink creamy milk. Lisa is excited because on Saturday, Whiskers turns two years old. After school on Friday, Lisa rushes to the pet store. She wants to buy Whiskers' birthday presents. Last year, she gave Whiskers a play mouse and a blue feather. For this birthday, Lisa is going to give Whiskers a red ball of yarn and a bowl with a picture of a cat on the side. The picture is of a black cat. It looks a lot like Whiskers. What does Whiskers like to do? A. Sleep in the sun and drink creamy milk B. Play C. Drink D. Sleep

Table 12: Sample 2 of MMLU Dataset: Text

E Table Understanding Dataset

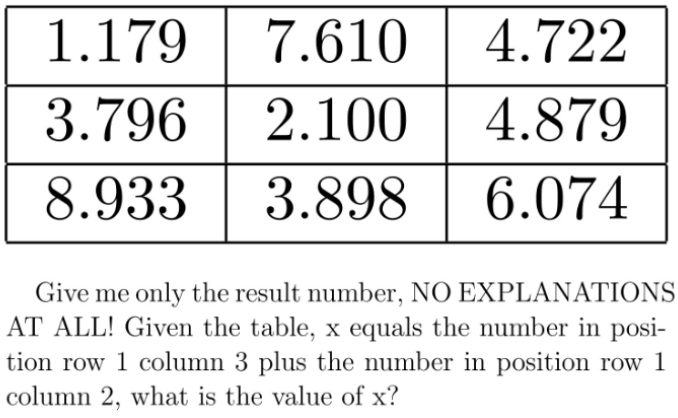

Figure 13: Sample 1 of Table Understanding Dataset: Image.

Text:	Give me only the result number, NO EXPLANATIONS AT ALL! Given the table, x equals the number in position row 1 column 3 plus the number in position row 1 column 2, what is the value of x? \begin{table}[] \centering \resizebox{\textwidth}{!}{% \begin{tabular}{llllll} \hline 1.179 & 7.610 & 4.722 \\ \hline 3.796 & 2.100 & 4.879 \\ \hline 8.933 & 3.898 & 6.074 \\ \hline \end{tabular}} \end{table}

Table 13: Sample 1 of Table Understanding Dataset: Text

9.875	3.149	3.765	5.892	1.333
6.335	3.325	3.529	9.173	6.089
2.789	4.895	5.894	9.548	0.213
3.692	6.280	2.986	6.015	1.774
1.852	7.581	8.438	2.641	7.873

Give me only the result number, NO EXPLANATIONS AT ALL! Given the table, x equals the number in position row 5 column 3 plus the number in position row 1 column 4, what is the value of x?

Figure 14: Sample 2 of Table Understanding Dataset: Image.

Text: Give me only the result number, NO EXPLANATIONS AT ALL! Given the table, x equals the number in position row 5 column 3 plus the number in position row 1 column 4, what is the value of x?

```
\begin{table}[]
\centering
\resizebox{\textwidth}{!}{%
\begin{tabular}{|l|l|l|l|l|}
\hline
9.875 & 3.149 & 3.765 & 5.892 & 1.333
\\
\hline
6.335 & 3.325 & 3.529 & 9.173 & 6.089
\\
\hline
2.789 & 4.895 & 5.894 & 9.548 & 0.213
\\
\hline
3.692 & 6.280 & 2.986 & 6.015 & 1.774
\\
\hline
1.852 & 7.581 & 8.438 & 2.641 & 7.873
\\
\hline
\end{tabular}}
\end{table}
```

Table 14: Sample 2 of Table Understanding Dataset: Text

F Math Reasoning Dataset

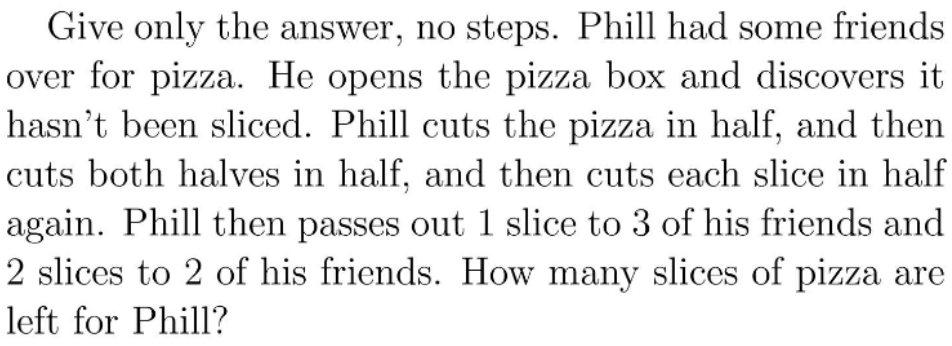

Figure 15: Sample 1 of Math Reasoning Dataset: Image.

Text:	Give only the answer, no steps. Phill had some friends over for pizza. He opens the pizza box and discovers it hasn't been sliced. Phill cuts the pizza in half, and then cuts both halves in half, and then cuts each slice in half again. Phill then passes out 1 slice to 3 of his friends and 2 slices to 2 of his friends. How many slices of pizza are left for Phill?

Table 15: Sample 1 of Math Reasoning Dataset: Text

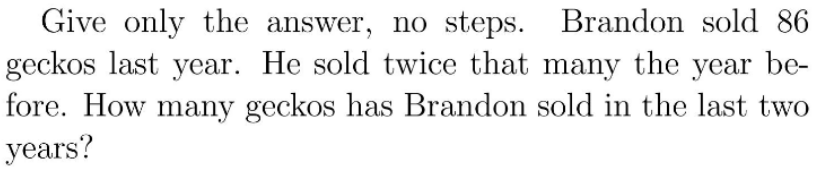

Figure 16: Sample 2 of Math Reasoning Dataset: Image.

Text:	Give only the answer, no steps. Brandon sold 86 geckos last year. He sold twice that many the year before. How many geckos has Brandon sold in the last two years?

Table 16: Sample 2 of Math Reasoning Dataset: Text

G State Machine Dataset

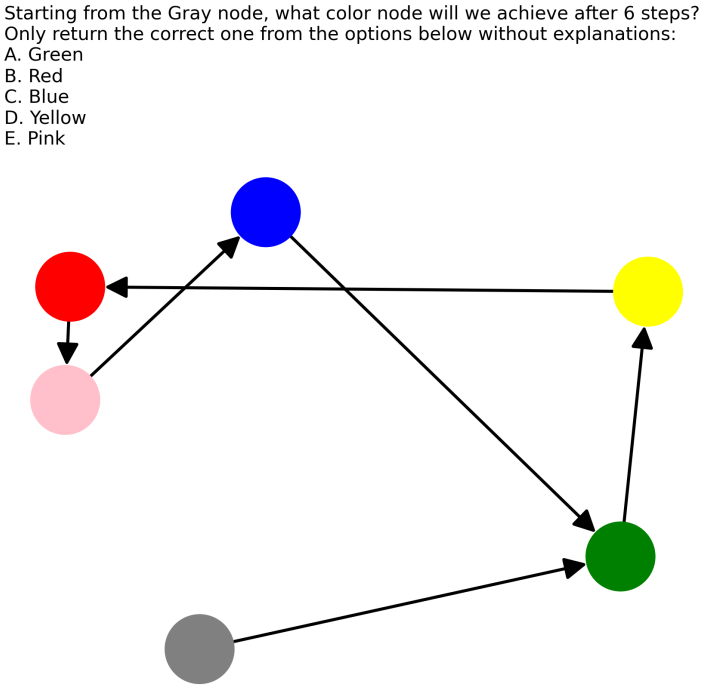

Figure 17: Sample 1 of State Machine Dataset: Image.

Text:	Consider a graph with the following directed edges: Yellow leads to Red; Green leads to Yellow; Red leads to Pink; Blue leads to Green; Gray leads to Green; Pink leads to Blue. Starting from the Gray node, what color node will we achieve after 6 steps? Only return the correct one from the options below without explanations: A. Green B. Red C. Blue D. Yellow E. Pink

Table 17: Sample 1 of State Machine Dataset: Text

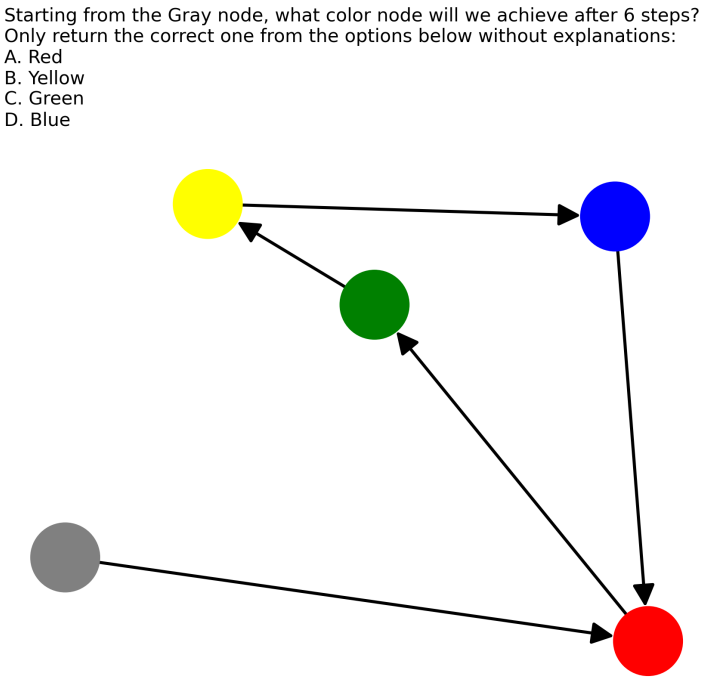

Figure 18: Sample 2 of State Machine Dataset: Image.

Text:	Consider a graph with the following directed edges: Gray leads to Red; Yellow leads to Blue; Blue leads to Red; Red leads to Green; Green leads to Yellow. Starting from the Gray node, what color node will we achieve after 6 steps? Only return the correct one from the options below without explanations: A. Red B. Yellow C. Green D. Blue

Table 18: Sample 2 of State Machine Dataset: Text