The Role of Chain-of-Thought in Complex Vision-Language Reasoning Task

Anonymous NAACL 2024 submission

Abstract

The study explores the effectiveness of the Chain-of-Thought approach, known for its proficiency in language tasks by breaking them down into sub-tasks and intermediate steps, in improving vision-language tasks that demand sophisticated perception and reasoning. We present the "Description then Decision" strategy, which is inspired by how humans process signals. This strategy significantly improves probing task performance by 50%, establishing the groundwork for future research on reasoning paradigms in complex vision-language tasks.

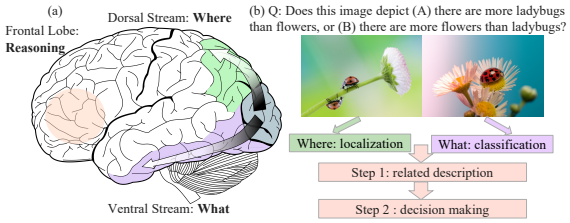


Figure 1: **Brain-inspired Two-step Reasoning.** Figure (a) is adapted from Wikipedia under license CC BY-SA 3.0 DEED. Figure (b) is adapted from Winoground (Thrush et al., 2022).

1 Introduction

Large language models (LLMs) have shown impressive performance in many language tasks, fostering the development of general AI assistants. An emerging trend in AI research aims to expand the potential of LLMs beyond text perception, incorporating visual data for more comprehensive models (Zhu et al., 2023; Liu et al., 2023; Jun et al., 2023). GPT-4V(ision) was recently released and has garnered significant interest for its exceptional abilities in multimodal perception and reasoning.

However, in complex vision-language tasks, although GPT-4V significantly outperforms the current state-of-the-art, it still lags behind human performance, as demonstrated in Section 3. We try to understand this challenge. Visual understanding extends beyond mere perception (Zellers et al., 2019). Complex visual-language tasks demand recognition-level perception, such as localizing and classifying objects and their attributes, as well as cognition-level reasoning, such as inferring intents, goals, and temporal and social dynamics. Humans can seamlessly integrate these two stages, but this remains challenging for LLMs with vision modality (vision-LLMs).

The Chain-of-Thought strategy, known for its effectiveness in language tasks by breaking them into sub-tasks with intermediate steps, has been studied extensively (Wei et al., 2022; Yao et al., 2023; Lyu et al., 2023), In this research, we investigate whether this method can enhance vision-language tasks, particularly those requiring complex reasoning. However, when confronting complex visiolinguistic tasks, current LLMs struggle to figure out the proper reasoning paradigm by themselves. Unlike math problems, which have a specific task-driven reasoning approach, these tasks may not have a clear and unique reasoning path leading to the final output.

Research on human cognition provides clues to a proper reasoning mode. Visual information propagates through two streams (Figure 1 (a)). The ventral stream (or the 'what pathway') is involved with object identification, while the dorsal stream (or the 'where pathway') processes objects' spatial location. These two streams decouple recognition into local processing modules. The cognition part, i.e., the reasoning and decision-making function, is mainly executed in the frontal lobe. Such modularization is a significant characteristic of brain structure and function (Gu et al., 2015; Bassett and Sporns, 2017). Inspired by how humans process signals, we design prompts with specialized modules for steps with different emphases.

The latest initial attempt at Chain-of-Thought on vision-language tasks focuses on the recognition task (Yang et al., 2023a), while another only pro-

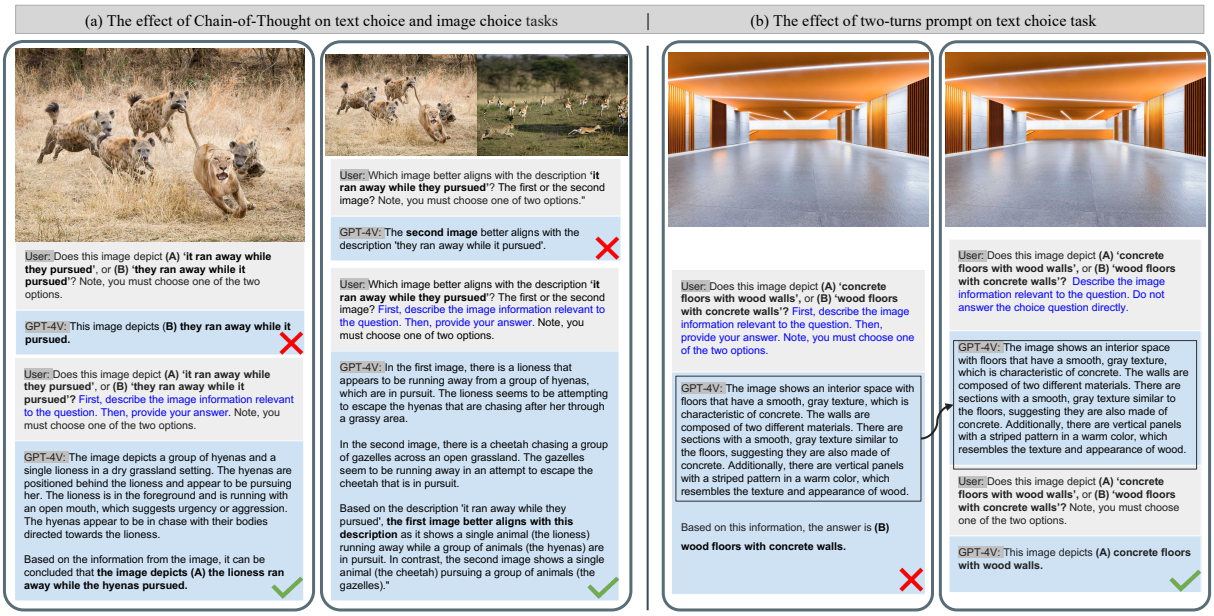


Figure 2: Examples of results with different prompt configurations. Text in blue highlights differences in the prompts. All figures shown here are from Winoground (Thrush et al., 2022)

vides qualitative analysis (Yang et al., 2023b). In this work, we analyze the prompting strategy for complex vision-language tasks analogous to the brain's information processing. Our *Description (information-extracting) then Decision (decision-making)* strategy consistently improves the performance across different experimental settings.

2 Probing Task

The ideal probing task should present sufficient challenges in both visual recognition and text comprehension. Furthermore, it should demand intricate reasoning, implying that there is a logical connection between the image and text facts leading to the final output. Consequently, we selected Winoground (Thrush et al., 2022) as a case study for our experiments. Winoground is both a dataset and a task designed specifically to evaluate visio-linguistic compositional reasoning. This task involves being given two images and two captions, with the goal of correctly matching each image with its corresponding caption. Notably, both captions utilize the exact same set of words, but they are arranged in different orders so that each caption primarily describes one of the two images.

The Winoground task is notably challenging (Diwan et al., 2022). It requires not only a robust visual recognition ability to identify small or blurred objects and differentiate attributes and actions among proximate objects but also sophisticated visio-linguistic compositional reasoning. For instance, solving Winoground sometimes necessitates interpreting images non-literally due to the idiomatic language usage in a caption (e.g., "it starts with Z and ends with A" might describe an image of a zebra). In other instances, one might need to infer past or future events from the scenes depicted in the image (e.g., "the cup on the left was filled first, and the cup on the right was filled second"). Examples of Winoground can be found in Fig 2.

The original Winoground task consists of two experimental setups: the text score and the image score. The text score evaluates the model's ability to select the correct caption from two given captions when provided an image. Conversely, the image score assesses the model's ability to choose the appropriate image from two available options given a caption. The original Winoground was tested based on the feature embedding similarities between captions and images in vision-language models, such as CLIP (Radford et al., 2021). To assess recent large vision-language models like GPT-4V, we reformulate the Winoground as a choice-based visual question-answering task. Formally, given images I_0 and I_1 and captions C_0 and C_1, the text score for a data point (C_0, I_0, C_1, I_1) is computed as follows:

$$s\left(C_0, I_0, C_1, I_1\right) = \begin{cases} 1 & \text{if } f\left(C_0, C_1, I_0\right) = C_0, \\ & \text{and } f\left(C_0, C_1, I_1\right) = C_1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

, where $f(\cdot)$ is the large language model that pro-

vides answers through a generation process. For a data point to be classified as correct, both images in a pair must align with their textual descriptions.

Similarly, for the image choice task, the score is determined as follows:

$$s\left(C_0, I_0, C_1, I_1\right) = \begin{cases} 1 & \text{if } f\left(I_0, I_1, C_0\right) = I_0, \\ & \text{and } f\left(I_0, I_1, C_1\right) = I_1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

The Winoground dataset comprises 400 pairs of images and their corresponding captions.

3 Evaluation

We first present the Winoground benchmark in Table 1 to evaluate the performance of GPT-4V in comparison to others. Note that while the original task evaluation is based on the encoding similarities between images and text, the setups for large vision-language models may vary slightly among methods. For example, TIFA (Hu et al., 2023) and VQ2 (Yarom et al., 2023) ask a series of questions given one image and one caption, then accumulate the scores. MMICL (Zhao et al., 2023) is given two images and two captions in each prompt. While in our set-up, we are given one image with two captions, or two images with one caption, as formulated in Eqn 1 and Eqn 2. However, these variations will not impact the primary objective of our study, which is to analyze the effects of various prompt configurations.

3.1 The Role of Chain-of-Thought

We analyze whether using a chain-of-thought prompt strategy, by decomposing the visual-language complex reasoning task into recognition and reasoning steps, can be beneficial compared to directly asking the model for the answer. The quantitative results are reported in Table 1, and the qualitative examples are shown in Figure 2 (a), with the prompt configurations detailed accordingly. In the following, we present the prompt we use to evaluate GPT-4V's text score (*Text*) and image score (*Image*), without and with CoT.

GPT-4V (Text): ['image-0' or 'image-1'] Does this image depict (A) ['caption-0'], or (B) ['caption-1']? Note, you must choose one of the two options.

GPT-4V CoT (Text): ['image-0' or 'image-1'] Does this image depict (A) ['caption-0'], or (B) ['caption-1']? First, describe the image information relevant to the question. Then, provide your answer. Note you must choose one of the two options.

Model & Prompt	Text	Image	Group
Random Chance	25.00	25.00	16.67
MTurk Human	89.50	88.50	85.50
CLIP-based encoding similarity			
CLIP (Radford et al., 2021)	30.75	10.50	8.00
METER (Dou et al., 2022)	44.99	22.75	18.75
Fiber (Wang et al., 2023)	51.49	31.49	27.50
Vision Large Language Models			
TIFA (Hu et al., 2023)	19.00	12.50	11.30
PALI (Chen et al., 2023)	46.50	38.00	28.75
VQ2 (Yarom et al., 2023)	47.00	42.20	30.50
MMICL (Zhao et al., 2023)	45.50	44.99	43.00
GPT-4V	69.25	46.25	39.25
GPT-4V CoT	**75.25**	**68.75**	**58.75**

Table 1: **Results on the Winoground task.** The scores (%) for various models are reported in the following references: MTurk Human and CLIP in (Thrush et al., 2022), METER and Fiber in (Wang et al., 2023), TIFA, PALI, and VQ2 in (Yarom et al., 2023), and MMICL in (Zhao et al., 2023). Note that the CLIP-based encoding similarity scores are derived from a deterministic process, while the results for the large vision-language models are obtained through a generative process.

GPT-4V (Image): ['image-0'], ['image-1'] Which image better aligns with the description ['caption-0' or 'caption-1']? The first image or the second image? Note you must choose one of two options.

GPT-4V CoT (Image): ['image-0'], ['image-1'] Which image better aligns with the description ['caption-0' or 'caption-1']? The first image or the second image? First, describe the image information relevant to the question. Then, provide your answer. Note you must choose one of two options.

As shown in Table 1, there are consistent improvements in both the *Text* and *Image* score settings, leading to a 50% improvement (from 39.25 to 58.75) in the *Group* score. The *Group* score represents the percentage of data points that have both *Text* and *Image* probes answered correctly. The "Description then Decision" strategy is particularly beneficial for the *Image* score setting, which improves from 46.25 to 68.75. We observe that for GPT-4V, the image score is significantly lower than the text score. This could be attributed to the use of multiple images as input. The Chain of Thought (CoT) strategy significantly improves the image score by generating task-relevant image descriptions and largely closes the gap with the text score.

The difference in generative processes between these two prompts is as follows: for GPT-4V, the answer is conditioned on the image and question:

$$\text{P(Answer | Image, Question)} \quad (3)$$

Model & Prompt	Text
(a) The Effect of CoT on Other Vision-LLMs	
InstructBLIP (Dai et al., 2023)	17.50
InstructBLIP CoT	31.50
LLaVA (Liu et al., 2023)	25.00
LLaVA CoT	33.50
(b) The Effect of Two-Turns prompt	
GPT-4V (1-turn)	69.25
GPT-4V CoT (1-turn)	75.25
GPT-4V Desp + GPT-4 QA (2-turns)	72.25
GPT-4V Desp + GPT-4 CoT (2-turns)	78.75
GPT-4V Desp + GPT-4V QA (2-turns)	79.50
GPT-4V Desp + GPT-4V CoT (2-turns)	80.00

Table 2: **Model Performance Comparison.** Section (a) presents performance comparisons of vision large language models, specifically InstructBLIP and LLaVA, with and without the utilization of Chain-of-Thought (CoT) prompting. Section (b) examines the effects of employing two-turn prompts on GPT-4V performance.

While for GPT-4V CoT, the answer is generated based on the image, question, and description:

$$\text{P(Answer | Image, Question, Description)} \quad (4)$$

Although generating descriptions from images does not introduce new information, our results show that this step simplifies the reasoning or decision-making for the model by translating visual signals into textual ones. To assess generalizability, we conducted the same experiments on other vision-based large language models, LLaVA (Liu et al., 2023) and InstructBLIP (Dai et al., 2023), as shown in Table 2 (a). The "description then decision" strategy consistently improves performance. More qualitative results are shown in the appendix.

3.2 The Effect of Two-turns Prompt

While examining the output of "GPT-4V CoT", we observed instances where correct descriptions were followed by incorrect answers. Therefore, to simplify the generation, we conducted experiments to divide the recognition and reasoning into two turns, on the *text* setting, as illustrated in Fig. 2 (b).

To further examine the quality and role of the first turn, i.e., the question-relevant image description, we conducted ablation studies employing GPT-4 in the second turn, which generates responses without image access. Another slight variation in the second turn's prompt includes adding the directive, "First, analyze the two options, then make your choice". In the following, we present the prompts we used for different settings of the second turn.

GPT-4 QA: ['image description'] Based on this image description, does this image depict (A) ['caption-0'], or (B) ['caption-1']? Note, you must choose one of the two options.
GPT-4 CoT: ['image description'] Based on this image description, does this image depict (A) ['caption-0'], or (B) ['caption-1']? First, analyze the two options, then make your choice. Note, you must choose one of the two options.
GPT-4V QA: ['image description'] Does this image depict (A) ['caption-0'], or (B) ['caption-1']? Note, you must choose one of the two options.
GPT-4V CoT: ['image description'] Does this image depict (A) ['caption-0'], or (B) ['caption-1']? First, analyze the two options, then make your choice. Note, you must choose one of the two options.

The results are summarized in Table 2 (b). Our observations are threefold: 1) The two-turn prompt notably improves results, from 75.25% to 79.50%. 2) The QA performance of GPT-4 reflects the quality of the image descriptions generated by GPT-4V. The "GPT-4V Desp + GPT-4 CoT (2-turns)" experiment achieved a 78.75% performance rate, validating the preciseness of GPT-4V's image descriptions. 3) GPT-4V performs slightly better than GPT-4, indicating that while the text format makes reasoning easier, fully enumerating all related information presented in an image remains challenging.

4 Conclusion and Limitation

In this study, we introduce a "description then decision" strategy for vision-language tasks. From a neuroscience perspective, humans conduct recognition and reasoning in distinct modules and through multiple steps. From a model training perspective, large language models are proficiently trained on linguistic tasks, and vision encoders have increasingly been aligned with these language models through image captioning. Given a vision-language task, the "description then decision" approach transforms the task into two well-trained tasks. Although straightforward, our prompt strategy has demonstrated consistent improvements across various models, paving the way for future research into reasoning paradigms for vision-language tasks.

A limitation of this study is that for simpler tasks, the benefits of this prompt strategy may not be fully reflected in the quantitative metrics. Future research could explore the scalability of this approach across a spectrum of task difficulties and its applicability in different domains.

References

Danielle S Bassett and Olaf Sporns. 2017. Network neuroscience. *Nature neuroscience*, 20(3):353–364.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2023. Pali: A jointly-scaled multilingual language-image model. *ICLR*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*.

Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. 2022. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *EMNLP*.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. 2022. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*.

Shi Gu, Theodore D Satterthwaite, John D Medaglia, Muzhi Yang, Raquel E Gur, Ruben C Gur, and Danielle S Bassett. 2015. Emergence of system roles in normative neurodevelopment. *Proceedings of the National Academy of Sciences*, 112(44):13681–13686.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2023. Equivariant similarity for vision-language foundation models. In *ICCV*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. 2023a. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of Thoughts: Deliberate problem solving with large language models. In *NeurIPS*.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2023. What you see is what you read? improving text-image alignment evaluation. *arXiv preprint arXiv:2305.10400*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *CVPR*.

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. 2023. Mmicl: Empowering vision-language model with multi-modal in-context learning. *arXiv preprint arXiv:2309.07915*.

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

A Appendix

A.1 Error Analysis

The performance of GPT-4V on the Winoground is remarkable, yet it's essential to pinpoint its limitations for a comprehensive understanding. Our error analysis, which adapts Winoground's tagging categorization, examines GPT-4V's performance on vision-language understanding tasks that require various recognition and reasoning skills. The tagging protocol is detailed in Table 4.

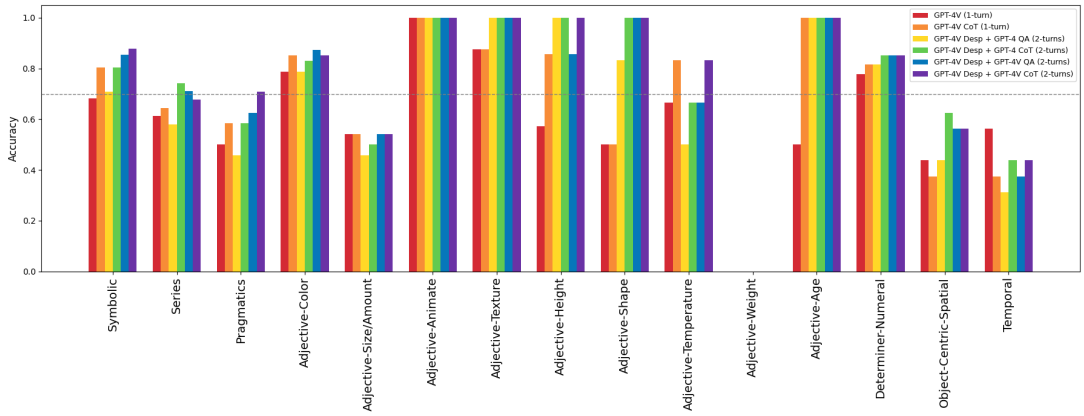


Figure 3: **Error analysis by tag category across different GPT-4V prompt configurations.** Each bar represents the accuracy of a specific experiment configuration on a given tag category. The gray dashed line denotes a baseline score of 0.7. From left to right, the experiments are: GPT-4V (1-turn), GPT-4V CoT (1-turn), GPT-4V Desp + GPT-4 QA (2-turns), GPT-4V Desp + GPT-4 CoT (2-turns), GPT-4V Desp + GPT-4V QA (2-turns), and GPT-4V Desp + GPT-4V CoT (2-turns).

Tag	Correct|Tagged	Accuracy
Across-Data-Set	[320 | 400]	(89.50%) 80.00%
Symbolic	[36 | 41]	(96.46%) 87.80%
Series	[21 | 31]	(96.65%) 67.74%
Pragmatics	[17 | 24]	(58.82%) 70.83%
Adjective-Color	[40 | 47]	85.11%
Adjective-Size/Amount	[13 | 24]	54.17%
Adjective-Animate	[9 | 9]	100.00%
Adjective-Texture	[8 | 8]	100.00%
Adjective-Height	[7 | 7]	100.00%
Adjective-Shape	[6 | 6]	100.00%
Adjective-Temperature	[5 | 6]	83.33%
Adjective-Weight	[0 | 3]	0.00%
Adjective-Age	[2 | 2]	100.00%
Determiner-Numeral	[23 | 27]	85.19%
Object-Centric-Spatial	[9 | 16]	56.25%
Temporal Dynamics	[7 | 16]	43.75%

Table 3: **Error Analysis by Tag Category.** The table presents the number of correct items versus the total items tagged in each category, alongside the corresponding accuracy. The results reported here are from the experiment "GPT-4V Desp + GPT-4V CoT", with an overall accuracy of 80% on text score. The percentages in parentheses present the MTurk Human performance.

The experiment selected for this analysis is "GPT-4V Desp + GPT-4V CoT (2-turns)," which achieves 80% text score across the dataset. As per Table 3, GPT-4V is adept at interpreting symbolic categories, which represent symbolic representations, e.g., a child's drawing. The "Series" tag indicates that a pair of images comes from the same photographic series, which might include identical individuals and scenes. While humans quickly understand semantic differences, visual similarities can pose challenges to the model, leading to lower accuracy compared to the baseline. The "Pragmatics" tag is for images that require non-literal interpretation, such as understanding idiomatic language or clarifying syntactic ambiguities in captions. Challenges such as discerning metaphors or the specific semantic relationships in prepositional phrases are reflected in the lower accuracy rates for both humans and models compared to the baseline.

In differentiating attributes, objective ones based on visual facts, such as color and shape, are easier to discern. In contrast, more abstract attributes, like size or amount, which require a reference, or weight, which necessitates additional knowledge, are more challenging to identify.

The "Determiner Numeral" category, which usually involves counting, is handled competently. Object-centric spatial tasks require identifying the correct reference frame and understanding the spatial relationships from an object's perspective in the image, which demands an interpretation of the 3D world from a 2D representation; this remains challenging and is generally more difficult than discerning camera-view spatial relationships. Temporal dynamics, which involve inferring past, present, and future events, are still difficult for the model.

We present the error analysis in Fig. 3 for all experimental settings shown in Table 2. The results consistently show that the categories tagged with 'Series, Pragmatics, Size/Amount, Weight, Object-Centric Spatial, Temporal' are comparatively more difficult. Error examples are presented in Fig. 6.

Tag	IDs
Symbolic	31, 32, 41, 50, 54, 55, 61, 62, 70, 78, 83, 89, 93, 107, 110, 136, 164, 174, 202, 203, 208, 209, 217, 219, 229, 238, 245, 246, 247, 254, 267, 274, 275, 303, 314, 320, 327, 339, 372, 383, 393
Series	2, 10, 23, 27, 41, 50, 67, 70, 87, 201, 291, 316, 335, 336, 338, 340, 341, 344, 345, 354, 355, 356, 357, 358, 367, 369, 372, 380, 383, 385, 388
Pragmatics	75, 76, 77, 78, 88, 180, 218, 219, 221, 222, 229, 262, 263, 267, 275, 307, 310, 323, 343, 344, 348, 353, 355, 394
Adjective-Color	28, 29, 30, 32, 67, 70, 71, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 145, 156, 199, 203, 208, 225, 232, 238, 252, 253, 259, 260, 261, 266, 269, 273, 290, 291, 304, 311, 312, 330, 335, 341, 342, 358, 359, 391
Adjective-Size/Amount	18, 19, 20, 40, 42, 43, 65, 80, 111, 119, 120, 121, 122, 145, 146, 170, 185, 187, 206, 209, 213, 280, 299, 307
Adjective-Animate	38, 39, 40, 41, 100, 101, 102, 103, 277
Adjective-Texture	36, 37, 82, 125, 126, 127, 283, 346
Adjective-Height	1, 26, 66, 157, 200, 223, 308
Adjective-Shape	33, 34, 35, 117, 118, 282
Adjective-Temperature	45, 124, 281, 295, 298, 303
Adjective-Weight	128, 206, 299
Adjective-Age	0, 207
Determiner-Numeral	6, 14, 59, 60, 61, 62, 63, 69, 93, 134, 135, 155, 163, 197, 204, 205, 227, 287, 317, 318, 350, 370, 379, 380, 383, 386, 390
Object-Centric-Spatial	34, 67, 81, 201, 217, 225, 226, 230, 231, 268, 328, 338, 347, 357, 378, 397
Temporal Dynamics	46, 57, 58, 130, 176, 189, 190, 191, 192, 193, 249, 311, 312, 316, 330, 331

Table 4: Tagged data IDs by category in Winoground as used in Table 3.

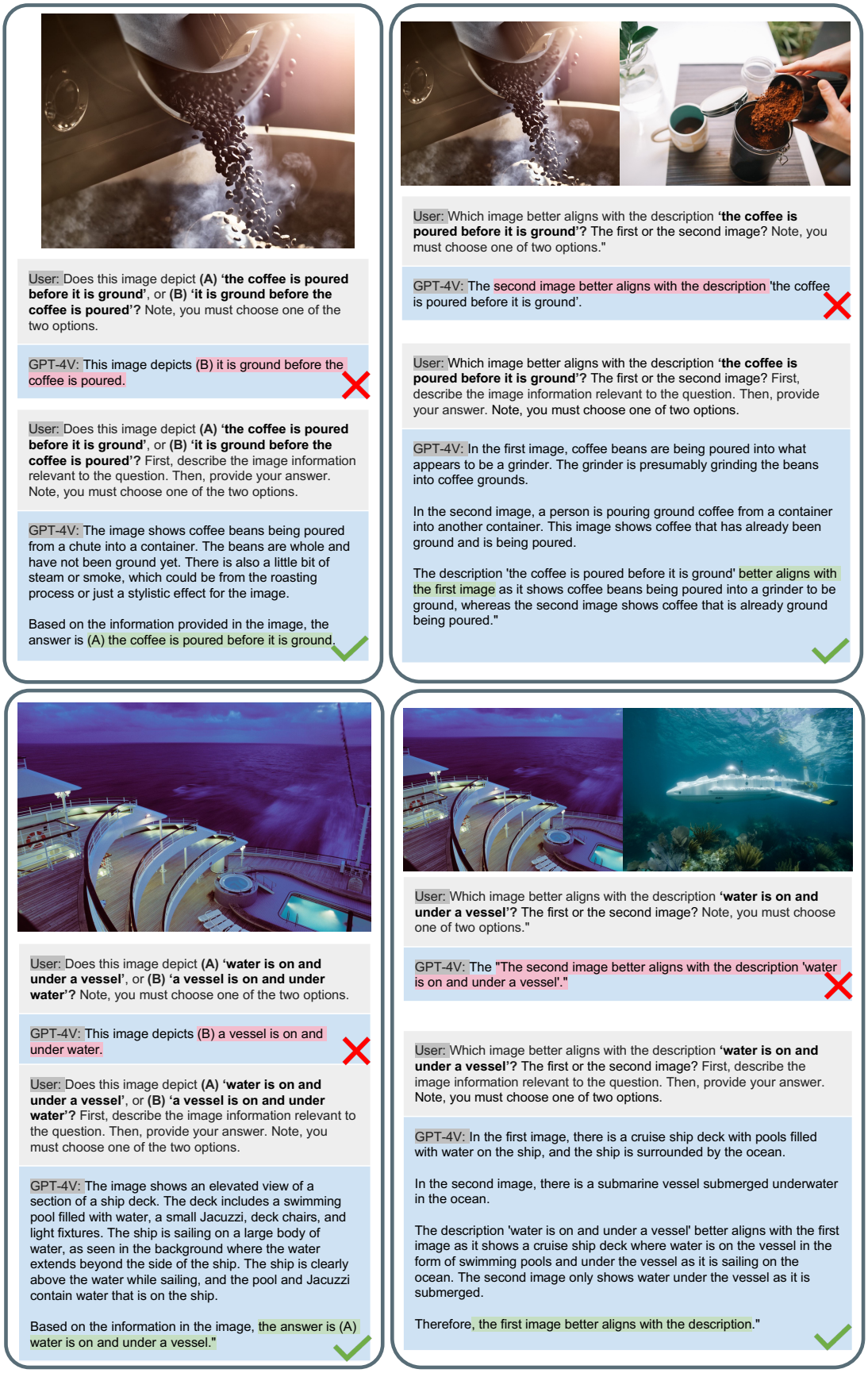


Figure 4: The qualitative examples demonstrate the effect of Chain-of-Thought. For each example, the left side illustrates the *text* choice setting and the right side depicts the *image* choice setting. The top portion shows the outcome without Chain-of-Thought, referred to as "GPT-4V (1-turn)", while the bottom part shows the results with Chain-of-Thought, labeled as "GPT-4V CoT (1-turn)" in Table 1. All images shown here are from Winoground (Thrush et al., 2022).

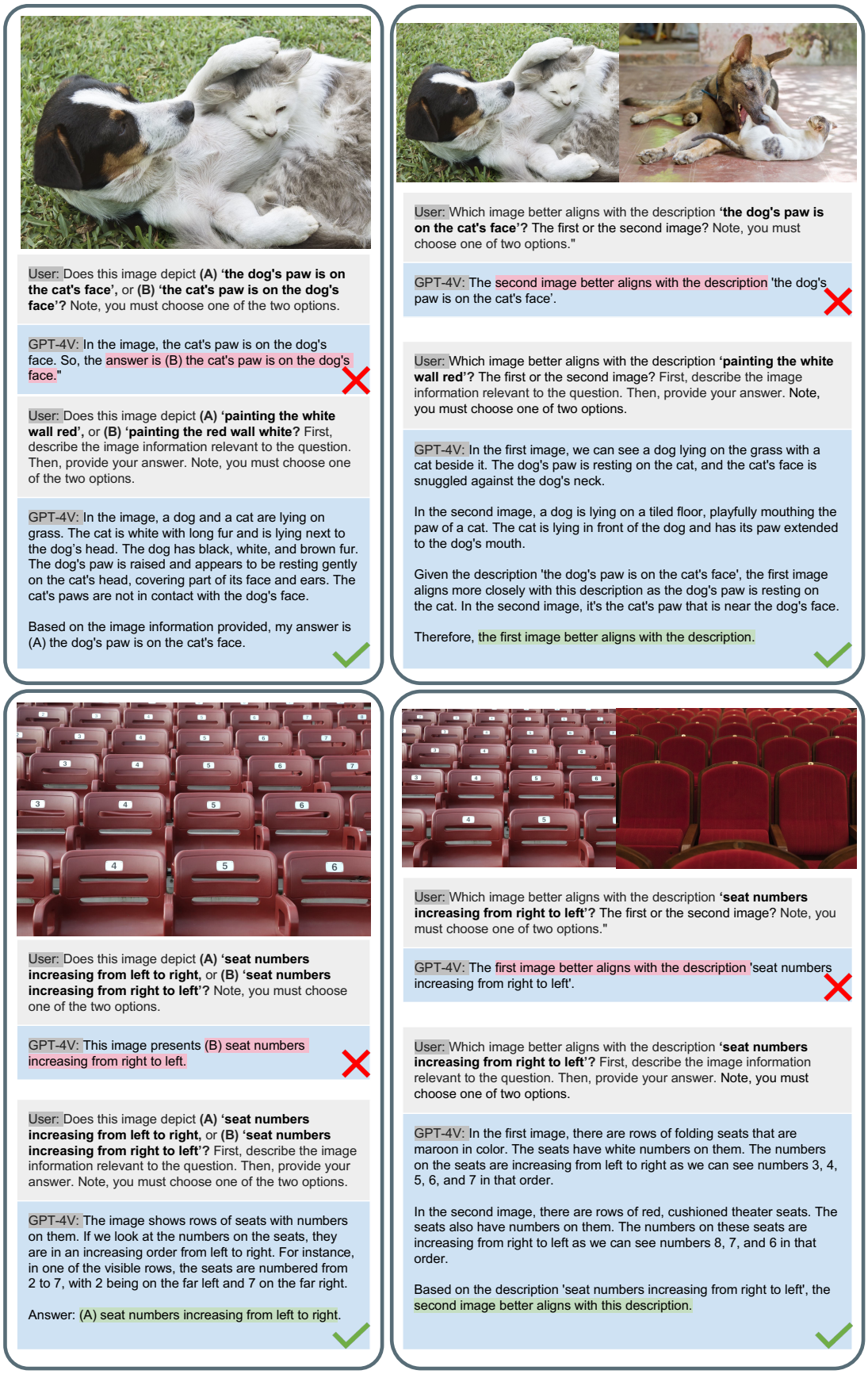


Figure 5: The qualitative examples (continued) demonstrate the effect of Chain-of-Thought. For each example, the left side illustrates the *text* choice setting and the right side depicts the *image* choice setting. The top portion shows the outcome without Chain-of-Thought, referred to as "GPT-4V (1-turn)", while the bottom part shows the results with Chain-of-Thought, labeled as "GPT-4V CoT (1-turn)"" in Table 1. All images shown here are from Winoground (Thrush et al., 2022).

User: Does this image depict **(A) "there's more milk than coffee",** or **(B) there's more coffee than milk'?** Describe the image information relevant to the question. Do not answer the choice question directly.

GPT-4V: The image depicts a glass mug that is filled with a beverage. The bottom part of the beverage in the mug is white, which is consistent with the appearance of milk. Above the white layer, there is a brown layer which is consistent with the appearance of coffee. The brown layer occupies more than half of the mug, while the white layer occupies less than half.

User: Does this image depict **(A) "there's more milk than coffee",** or **(B) there's more coffee than milk'?** First, analyze the two options, then make your choice. Note, you must choose one of the two options.

GPT-4V: Analyzing the two options:
(A) There's more milk than coffee - This would mean that the white layer at the bottom of the mug occupies more space than the brown layer.
(B) There's more coffee than milk - This would mean that the brown layer in the mug occupies more space than the white layer.
Based on the image information where the brown layer occupies more than half of the mug and the white layer occupies less than half, I would choose option (B) There's more coffee than milk.

Tag: Adjective-Size/Amount

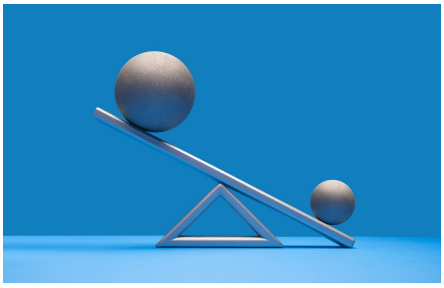
User: Does this image **(A) 'the cup on the left is filled first and the cup on the right is filled second',** or **(B) 'the cup on the left is filled second and the cup on the right is filled first'?** Describe the image information relevant to the question. Do not answer the choice question directly.

GPT-4V: The image shows a person pouring a steaming, amber-colored liquid from a clear round teapot into a blue cup on the left. There is another blue cup on the right which is not being filled, and it does not have any visible steam coming from it. The teapot contains slices of citrus and some leaves, possibly herbs. The focus is on the left cup being filled, and the pouring liquid is creating a stream from the teapot into the cup.

User: Does this image **(A) 'the cup on the left is filled first and the cup on the right is filled second',** or **(B) 'the cup on the left is filled second and the cup on the right is filled first'?** First, analyze the two options, then make your choice. Note, you must choose one of the two options.

GPT-4V: Analyzing the two options:
(A) The cup on the left is filled first and the cup on the right is filled second: This option suggests that the person is pouring the liquid into the left cup first, and will pour into the right cup next.
(B) The cup on the left is filled second and the cup on the right is filled first: This option suggests that the person has already poured liquid into the right cup and is now pouring into the left cup.
Based on the information in the image, where the left cup is being filled with a steaming liquid and the right cup is not being filled and shows no signs of steam, I would choose option (A) The cup on the left is filled first and the cup on the right is filled second.

Tag: Temporal Dynamics



User: Does this image depict **(A) "a small spherical object is heavier than a big spherical object",** or **(B) a big spherical object is heavier than a small spherical object'?** Describe the image information relevant to the question. Do not answer the choice question directly.

GPT-4V: The image shows a large sphere on one end of a seesaw and a smaller sphere on the other end. The seesaw is tilted towards the side where the larger sphere is placed, indicating that it is weighing down that side. The smaller sphere is on the elevated side of the seesaw, suggesting that it is lighter.

User: Does this image depict **(A) "a small spherical object is heavier than a big spherical object",** or **(B) a big spherical object is heavier than a small spherical object'?** First, analyze the two options, then make your choice. Note, you must choose one of the two options.

GPT-4V: Option (A) suggests that the small spherical object is heavier than the big spherical object. Option (B) suggests that the big spherical object is heavier than the small spherical object.

Based on the image where the seesaw is tilted towards the side where the larger sphere is placed, indicating that it is weighing down that side, and the smaller sphere is on the elevated side of the seesaw, suggesting that it is lighter, I would choose option (B) - a big spherical object is heavier than a small spherical object."

Tag: Adjective-Size/Amount; Adjective-Weight

User: Does this image **(A) 'a more lightly-colored dog is holding a toy next to a black dog',** or **(B) 'a black dog is holding a toy next to a more lightly-colored dog'?** Describe the image information relevant to the question. Do not answer the choice question directly.

GPT-4V: The image shows two dogs; one black and one more lightly-colored (white/cream). They are both on a grassy surface. The black dog is to the left and the more lightly-colored dog is to the right. They both have their mouths on a colorful rope toy, which is predominantly red and yellow.

User: Does this image **(A) 'a more lightly-colored dog is holding a toy next to a black dog',** or **(B) 'a black dog is holding a toy next to a more lightly-colored dog'?** First, analyze the two options, then make your choice. Note, you must choose one of the two options.

GPT-4V: Analyzing the two options:
(A) more lightly-colored dog is holding a toy next to a black dog.
(B) A black dog is holding a toy next to a more lightly-colored dog.

Based on the image information provided earlier, the correct choice is (B) a black dog is holding a toy next to a more lightly-colored dog.

Tag: Adjective-Color

Figure 6: Examples of error cases with experiment "GPT-4V Desp + GPT-4V CoT (2-turns)", along with the category tags. All images shown here are from Winoground (Thrush et al., 2022).