RPTS-Eval: Evaluating Large Vision-Language Models for Reasoning Process and Scoring with Tree

Anonymous ACL submission

Abstract

Large Vision-Language Models (LVLMs) excel in multimodal reasoning, and have shown impressive performance across various multimodal benchmarks. However, most of these benchmarks evaluate models primarily through multiple-choice or short-answer formats, which do not take the reasoning process into account. Although some benchmarks do assess the reasoning process, their methods are often too simplistic and only examine reasoning when answers are incorrect. This approach overlooks scenarios where flawed reasoning leads to correct answers. In addition, these benchmarks do not consider the impact of inter-modal relationships on reasoning. To address this issue, we propose RPTS-Eval, a benchmark focused on meticulously evaluating the reasoning process of models. RPTS-Eval comprises 374 images and 390 reasoning instances, covering 6 types of vision-language capabilities. We also introduce a new evaluation metric called RPTS to provide a fine-grained reflection of the reasoning process, which can not only indicate the overall correctness of the reasoning but also pinpoint the specific step where the model makes an error. We evaluated representative LVLMs (e.g., GPT-4o, Llava-Next), uncovering their limitations in multimodal reasoning and highlighting the differences between open-source and closed-source commercial LVLMs. We believe that this benchmark will contribute to advancing research in the field of multimodal reasoning.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable linguistic capabilities, particularly in the task of natural language inference, showing impressive performance(OpenAI et al., 2024). In the real world, information is obtained through various channels, including visual and auditory, not solely through language. This realization has led to the development of Large Vision-Language Models (LVLMs)(Li et al., 2023; Liu et al., 2024c; Alayrac et al., 2022), aimed at equipping models with advanced cognitive abilities. To enhance these models, it is crucial to assess their reasoning abilities, which will guide further improvements. However, the diversity of potential reasoning paths that lead to the same conclusion presents a significant challenge in evaluating the reasoning abilities of these models.

Previous research has predominantly employed high-quality, multi-capability(Yu et al., 2023), and challenging benchmarks(Yue et al., 2024) to evaluate the reasoning abilities of models. These studies have contributed significantly to the evaluation of model reasoning abilities. However, they often bypassed the complexity of the reasoning process itself, instead focusing on the final answers of the models through multiple choice and short answer formats. Only a few studies, such as InfiMM-Eval(Han et al., 2023), have incorporated the reasoning process into their evaluations, scoring it when the answers were incorrect. This strategy prevents misjudging correct answers due to different expressions, but fails when correct answers are derived from flawed reasoning, as show in Figure 1. Furthermore, InfiMM-Eval assesses the reasoning process by inputting it into a LLM, a method that does not allow for a finer analysis of the reasoning itself.

In the field of NLP, recent work has proposed several methods for validating the reasoning process ((Golovneva et al., 2023), (Prasad et al., 2023)). These methods primarily focus on ensuring the logical consistency of linear reasoning, where each step must not contradict the previous steps. However, such approaches are not well-suited for multimodal contexts. As shown in Figure 1, the information derived from an image may conflict with that from text, but both can still be correct and lead to the right answer. Furthermore, these methods do not integrate the evaluation of the reasoning

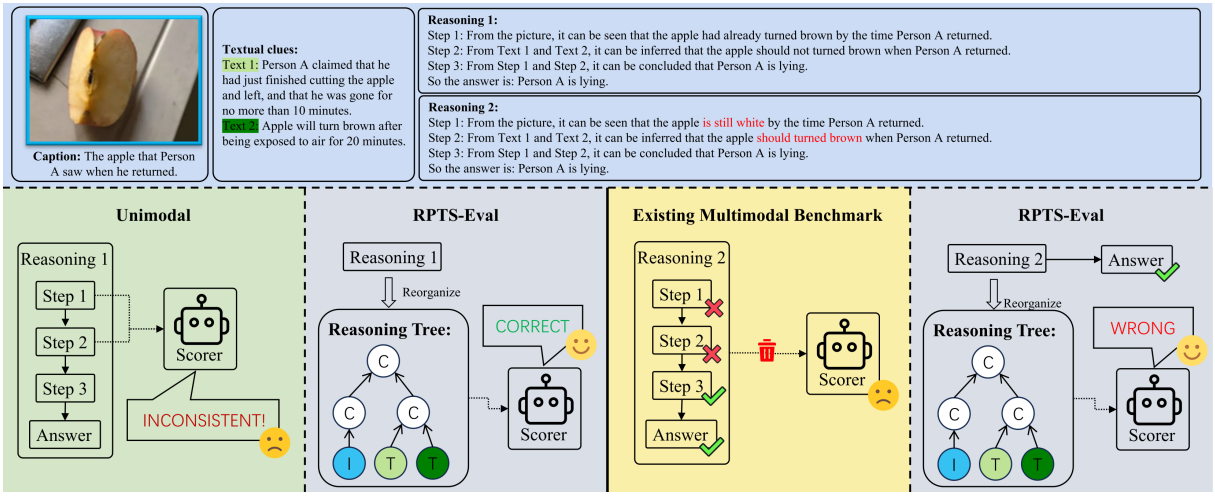

Figure 1: Comparison between Unimodal, existing Multimodal benchmarks and our RPTS-Eval. **Left**: The unimodal approach is unable to handle reasoning involving conflicting information across different modalities. **Right**: Current multimodal benchmarks fail to detect instances where reasoning errors are present, yet the answer remains correct.

process with the final metric, failing to intuitively reflect the impact of reasoning quality on the result.

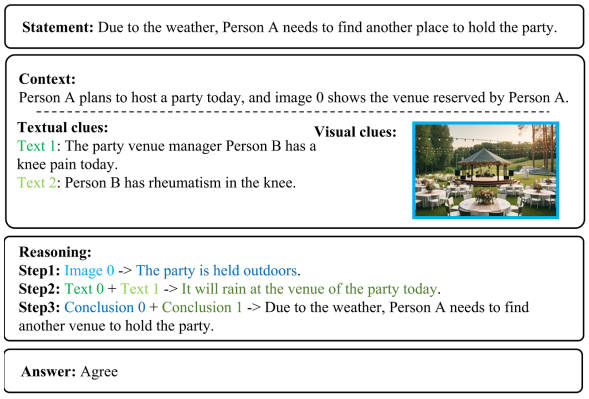

Figure 2: An example of RPTS-Eval.

To enhance the precision of model reasoning analysis, we developed RPTS-Eval, a bilingual benchmark featuring a specialized reasoning format. This benchmark comprises 374 images and 390 reasoning tasks, encompassing 6 visual language abilities such as image comparison, spatial awareness, and commonsense. We also define three types of relationships between modalities in the reasoning process: related without interference, related with interference, and unrelated. We utilized carefully crafted examples to facilitate GPT-4 in generating reasoning stories requiring diverse visual language abilities. Our annotators manually refined these narratives, organizing the reasoning into a specialized format, and selected appropriate images via the internet and text to image model. To ensure data quality, our staff manually annotated these reasoning and images, verifying the rigor of the reasoning logic. Figure 2 shows examples of RPTS-Eval. Each data in RPTS-Eval contains four parts: statement, context, visual clues, and textual clues. The model needs to infer whether the conclusions that can be drawn from visual and textual clues agree or disagree with the given statement.

The unique reasoning format of RPTS-Eval structures the reasoning process as a tree with visual and textual clues as leaf nodes and the conclusion as the non-leaf node, as illustrated in Figure 2. To address cases where the reasoning is flawed yet the answer is correct, we introduce a novel evaluation metric: the **R**easoning **P**rocess **T**ree **S**core (RPTS). The computation of RPTS relies on the reasoning tree and two parameters. By adjusting these parameters, RPTS can accurately assess the logic of the reasoning at both the global and local levels, thus enabling precise localization of reasoning errors. Experimental results demonstrate the efficacy of RPTS in these respects and highlight existing issues in open-source LVLMs. The primary contributions of our work can be summarized as follows:

- We develop RPTS-Eval, a new benchmark specifically designed for the multimodal reasoning domain. Compare with existing benchmarks, our benchmark focuses on the reasoning process and mandates a structured inference format, enabling a systematic evaluation of reasoning abilities.

- We define three types of relationships between modalities in reasoning, which clarify the classification of multimodal reasoning.
- We introduce a new metric, RPTS, for detecting correct conclusions based on faulty reasoning and genuinely logical reasoning processes, and reflecting both overall and local logic of reasoning, achieving error localization.
- We conduct extensive experiments on our RPTS-Eval. Results show that current open-source LVLMs struggle to extract conclusions for subsequent inference from images based on existing information, and demonstrate significant differences in model performance in different language contexts.

2 Related Work

2.1 MLLM Evaluation Benchmarks

Classic multimodal benchmarks typically evaluate specific reasoning abilities of models. For instance, OK-VQA(Marino et al., 2019) evaluates a model's capacity to leverage external knowledge for reasoning, while VCR(Zellers et al., 2019) focuses on human-related common sense reasoning. To assess a model's comprehensive abilities, researchers have proposed various benchmarks, such as MMBench(Liu et al., 2025), SEED-Bench(Li et al., 2024), MM-VET(Yu et al., 2023), and MMMU(Yue et al., 2024). These benchmarks scrutinize the reasoning abilities of models from diverse perspectives, often employing multiple-choice or simplified formats to facilitate the evaluation process. InfiMM-Eval(Han et al., 2023) incorporates the reasoning process into the evaluation, scoring the entire reasoning process. However, it cannot perform a more detailed analysis of the reasoning, and its evaluation method cannot exclude cases where incorrect reasoning leads to a correct answer.

2.2 Verify Reasoning Process

Recent studies have introduced various techniques for evaluating reasoning processes. ROSCOE(Golovneva et al., 2023) proposes a set of quality metrics to assess reasoning from four perspectives: semantic alignment, semantic similarity, logical correctness, and semantic coherence. ReCEval (Prasad et al., 2023) evaluates reasoning based on two criteria: whether the reasoning steps are correct and whether new information is derived from the reasoning. REVEAL provides a dataset to validate whether a model can be used to verify the reasoning process.

3 RPTS-Eval

3.1 Data Collection

We are aimed to developing a high-quality multimodal reasoning evaluation benchmark, using a meticulously designed methodology to assess model reasoning performance. Each sample in RPTS-Evalcan be viewed as a multimodal reasoning story. Constructing such stories automatically poses significant challenges, even GPT-4 struggles to generate reasoning stories with sufficiently coherent logic. In addition, it is difficult to find suitable stories from online sources, and the time investment required for manually designing stories is substantial. To address these issues, the process of constructing data can be broadly divided into the following steps:

Collating multimodal reasoning stories. To reduce the difficulty of manually designing stories, we use GPT-4 to assist annotators. First, we ask an annotator to design a few reasoning stories and input them into GPT-4 as examples. Following the approach of MM-Vet(Yu et al., 2023), we then require GPT-4 to generate reasoning stories encompassing six types of capabilities, based on the given examples. We serve these stories as starting point to lower the difficulty for annotators in designing reasoning tasks. The six capabilities are as follow:

- **Image Comparison(IC):** This involves the model comparing two images to find similarities or differences. This is a fundamental ability for humans, as we gain much information about the real world by comparing what we see.
- **Recognition(Rec):** Recognition refers to general visual capabilities, including identifying objects, object attributes, scenes, counting, and various other advanced visual recognition tasks in computer vision.
- **ORC:** Optical Character Recognition (OCR) involves understanding the text in images. The model needs to understand the text in images to complete subsequent reasoning tasks.
- **Spatial Awareness(SA):** Spatial awareness includes various spatial-related abilities, such

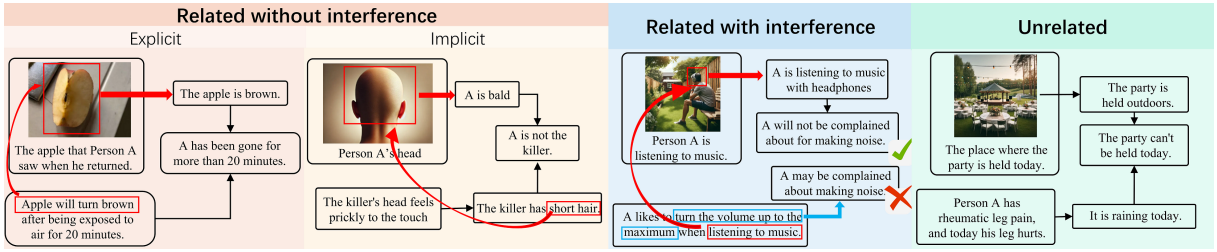

Figure 3: Three types of relationships between modalities. The red arrow means the relation between image and text, and the blue arrow means the interference between them.

as understanding absolute positional relationships from a fixed perspective and relative positional relationships that require perspective transformation.

- **Commonsense(Com):** Commonsense refers to general knowledge people have. In our daily decisions, not all information is presented to us, we need to use our existing knowledge to make decisions. For example, placing a glass cup in the middle of the table rather than on the edge. This ability requires the model not only to know these commonsense but also to select appropriate one based on specific scenarios to complete reasoning tasks.

- **Math:** Math ability assesses the model's capability to use arithmetic to aid in reasoning. For instance, if I know my best friend's monthly salary and his expenses for the month, I can deduce that he might need me to buy him a meal through simple calculations.

Constructing Data This phase involves two annotators, each assigned to different reasoning stories. First, the annotators need to design two reasoning paths based on the stories. These two reasoning paths should use similar clues to arrive at opposite conclusions. Then, the annotators should design statements, contexts, visual and textual clues, reasoning steps, and required abilities for the data based on the reasoning paths. Finally, the annotators need to find suitable images according to their design. The images for RPTS-Eval are sourced from the internet and text-to-image modals.

Quality Control To ensure data quality, each piece of data is validated by two validators. We reference InifMM-Eval(Han et al., 2023) and conduct a comprehensive evaluation of the data based on the following criteria:

- **Logical Scoring:** Carefully assess the relationship between statements, context, visual clues, and reasoning steps, and score them to ensure rigorous logic in the data.

- **Multimodality:** This criterion evaluates whether visual clues or textual clues are unnecessary for reasoning, filtering out samples that can be inferred using a single modality.

- **Subjectivity and Discrepancy Check:** If the problem is overly subjective or the validator's reasoning significantly differs from the ground truth, the data will be deleted or modified.

- **Missing or Redundant abilities:** Validators will judge, based on their reasoning experience, whether the annotated abilities are missing or redundant.

Multimodal Reasoning Classification To better investigate the reasoning capabilities of multimodal models, we categorize the constructed data into three types based on the relationships between modalities during reasoning. Examples of these three reasoning types are illustrated in Figure 3.

- **related without interference**: By utilizing information from one modality, it becomes possible to determine which information should be retrieved from another modality to complete the reasoning process. The relationships between modalities are categorized into two types: explicit and implicit. Explicit relationships are defined as cases where one modality directly indicates the information that needs to be obtained from another modality. In contrast, implicit relationships involve cues from one modality that require reasoning to infer which information should be retrieved from the other modality.

Statistics	Percentage	Statictic	Percentage
Capabilities			
Rec	83.08%	Math	24.87%
Com	40.00%	OCR	18.46%
SA	28.97%	IC	5.13%
Answer			
agree	50.00%	disagree	50.00%
Relationship			
rwoi	84.62%	rwi	6.92%
unrelated	8.46%		
Reasoning steps		Reasoning tree height	
≤ 2	3.85%	≤ 2	0.51%
3	42.82%	3	11.03%
4	32.56%	4	52.56%
5	13.08%	5	26.92%
≥ 6	7.69%	≥ 6	8.67%

Table 1: Key statistics of the RPTS-Eval benchmark. As each reasoning instance need one or more capabilities, the sum of percentage is larger than 100%. 'rwoi' and 'rwi' represent related without/with inference.

- **related with interference**: In addition to the aforementioned relationships, information from one modality can also mislead the extraction of information from another modality. This misguidance can manifest in two ways: either by extracting irrelevant or erroneous information from the other modality, or by failing to extract any information from it altogether.

- **unrelated**: The modalities are independent of each other, and information must be retrieved separately from each modality to complete the reasoning process.

Translation Finally, we use GPT-4 to translate the annotated Chinese data into English and make manual adjustments.

In summary, our RPTS-Eval benchmark comprises 390 inferences linked to a total of 374 images. Table 1 depicts the distribution across multiple dimensions of RPTS-Eval. Since most tasks require the recognition of objects in images, object recognition capability plays a dominant role. Given that the data is constructed with paired answers, the two types of answers in RPTS-Eval are evenly distributed, which helps mitigate the effects of model bias. The relationships between modalities are primarily based on related without interference, as the reasoning for the last two types are more challenging to construct. The majority of inferences can be made within 5 steps, and when the inference is represented as a tree, the tree height is typically below 6.

4 Experiments

4.1 Evaluation Protocol

Reorganize Model Output Our approach aims for more detailed evaluative reasoning by structuring the reasoning process as a tree. To construct this tree, the model is required to output reasoning in the "[PREMISE] + [PREMISE] -> [CONCLUSION]" format according to the RPTS-Eval annotation standard, where the '[PREMISE]' can be image cues, textual clues, or conclusions derived from previous steps. However, existing open-source LVLMs are not capable of strictly adhering to this format. To solve this issue, we first use a chain-of-thought prompt to guide the model in generating reasoning with premises step by step. Then, we employ GPT-4 to reformat the output of the LVLM into the required RPTS-Eval annotation format. Figure 4 illustrates the complete evaluation process.

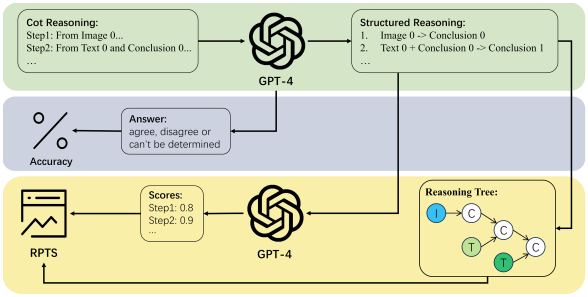

Figure 4: Evaluation process for RPTS-Eval

LLM-Based Scorer Now, each reasoning step in our approach strictly adheres to the "[PREMISE] + [PREMISE] -> [CONCLUSION]" format. Prior studies (Chiang and Lee, 2023; Liu et al., 2023; Fu et al., 2024; Bai et al., 2023b; Bitton et al., 2023; Yu et al., 2023; Han et al., 2023) have demonstrated GPT-4's effectiveness in assessing model reasoning. Therefore, we utilize GPT-4 to score reasoning, but with a unique twist: we only evaluate individual reasoning steps, not the entire process. This method allows for more precise evaluations by preventing the influence of other reasoning elements on the scores. Before we input the reasoning into scorer, we first preprocess the model's reasoning by eliminating redundant text clues, merging conclusions from images, substituting unnumbered texts and conclusions with all relevant clues and conclusions, and removing reasoning without '[PERMISE]'. For scoring reasoning according with image, we calcu-

late the semantic similarity of conclusions directly derived from images against the ground truth. For other reasoning, we input the premises and conclusion into GPT-4 to assess their logical coherence. The score given by scorer ranges from 0 to 1, with higher scores indicating stronger logical reasoning. However, as illustrated in Figure 1, there are instances where the model's selected premises may not directly support the given conclusion, though they may be justified within the broader reasoning context. To address this, if the initial score is below 0.5, we re-evaluate using all text clues and previously derived conclusions as new premises, and then applying a 0.8 penalty for incorrect premises. We select the higher of the two scores as the final assessment.

Reasoning Process Tree Score Considering the unique structure of reasoning, we can model the process as an reasoning tree, as depicted in figure 5. In this tree, the leaf nodes represent context, visual clues and textual clues, while the non-leaf nodes correspond to individual steps of inference. This tree, alongside parameters α and β, is used to weight each inferential step. The weight assigned to n_i is defined as

$$w_i = \alpha^{|\beta - h|} \tag{1}$$

where n_i is the node corresponding to the i^{th} step of inference, h denotes the height of n_i, defined as the number of edges on the longest path from n_i to any leaf node. The overall score of the reasoning tree, RPTS, is calculated as

$$RPTS = \frac{\sum_{i=1}^{N} w_i s_i}{\sum_{i=1}^{N} w_i} \tag{2}$$

where N is the number of steps in the inference process, and s_i is the score of the i^{th} inferential step. By adjusting α and β, we can finely tune the emphasis on global versus local aspects of the inference process. Figure 5 illustrates the scoring outcomes under three different settings of α and β. In the top, α is set to 1, making RPTS reflect the average score across all inferential steps. In the middle, α is 0.8 and β is 1, focusing RPTS more on the scores of earlier inferential steps. In the below, α is 0 and β is 2, meaning RPTS considers only the inference steps occurring at a tree height of 2, essentially focusing on inferences that derive directly from the clues.

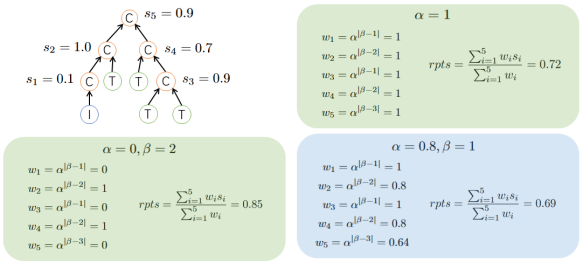

Figure 5: Examples of different parameter settings for RPTS. C, I and T respectively represent conclusion, visual clue and textual clue.

4.2 Models and Evaluation Metrics

To validate the challenging nature of RPTS-Eval and the capability of the RPTS evaluation metric analysis model, we conducted experiments in both Chinese and English across various models. The open-source models tested include InstructBLIP(Dai et al., 2024), Internval2(Chen et al., 2024), ShareGPT4V(Chen et al., 2023), Llava-v1.5(Liu et al., 2024a), Llava-Next(Liu et al., 2024b) and Qwen-VL-Chat(Bai et al., 2023a), detailed in Appendix A; the sole close-source model examined is GPT-4o. We evaluate the reasoning ability of the model by combining accuracy and RPTS, and analyze the problems of the model.

4.3 Experiment Settings

Our experiment involves both Chinese and English languages and performs chain-of-thought(COT)(Wei et al., 2022) reasoning on the RPTS-Eval benchmark. All tests were performed in a zero-shot setting using a greedy decoding strategy to assess the models' inferential abilities. To optimize the COT reasoning outcomes, we designed five Chinese prompts and seven English prompts, selecting the most effective one from each language for our experiments. All tests were carried out on an NVIDIA A100 GPU.

4.4 Results and Analysis

Table 2 displays the performance of various models on RPTS-Eval. In addition to evaluating the accuracy of the models' inferences and their mean RPTS scores, we applied an RPTS-based filter to exclude cases where incorrect inferences resulted in accurate conclusions. Specifically, we consider the reasoning logic with an RPTS score below 0.5 to be incoherent, and therefore classify it as incorrect. Appendix C provides two examples of inferences that were excluded under this criterion.

Models	English			Chinese		
	Acc	RPTS↑	$\text{Acc}_{filtered}$	Acc	RPTS↑	$\text{Acc}_{filtered}$
Llava-v1.5-7B	0.64	0.63	0.48(-0.16)	0.35	0.57	0.24(-0.12)
Llava-Next-7B	0.62	0.47	0.32(-0.29)	0.13	0.41	0.06(-0.07)
Qwen-VL-Chat	*0.57*	0.61	0.41(-0.16)	0.39	0.61	0.25(-0.14)
ShareGPT4V-7B	0.58	0.56	0.38(-0.20)	*0.34*	*0.50*	*0.19*(-0.15)
InternVL2-8B	0.63	0.67	0.53(-0.10)	0.46	0.66	0.37(-0.08)
Llama-3.2-11B	0.68	0.68	0.56(-0.12)	0.41	0.63	0.29(-0.12)
InstructBLIP	0.56	0.59	0.41(-0.16)	-	-	-
Llava-v1.5-13B	0.56	0.59	0.41(-0.15)	0.41	0.58	0.28(*-0.13*)
Llava-Next-13B	0.62	*0.51*	*0.34*(-0.27)	0.23	0.46	0.11(-0.12)
ShareGPT4V-13B	0.59	0.50	0.32(-0.27)	0.35	0.58	0.26(-0.09)
InternVL2-26B	0.65	0.70	0.55(-0.10)	0.54	0.74	0.45(-0.08)
Llava-Next-34B	0.68	0.71	0.60(-0.08)	0.46	0.68	0.37(-0.09)
InternVL2-40B	0.74	0.76	0.67 (-0.06)	0.57	0.75	0.52 (-0.05)
InternVL2-76B	0.73	0.79	0.70 (-0.04)	0.60	0.77	0.57 (-0.03)
Llama-3.2-90B	0.79	0.67	0.66(-0.12)	0.56	0.77	0.52 (-0.04)
GPT-4o	0.86	0.84	0.84 (-0.02)	0.72	0.86	0.70 (-0.02)

Table 2: Results of different models on RPTS-Eval with cot prompt. We set $\alpha = 0.9$, $\beta = 1$ when calculate RPTS. For each column, the highest, the second, and the third highest figures are highlighted by green, orange and pink backgrounds. The worst, second worst, and third worst are highlighted using underline, wavy underline, and *italic*, respectively. **Acc**: Accuracy.

The data in Table 2 reveal that all models exhibited a decline in accuracy to varying extents, with GPT-4o showing the least reduction. This modest decline is closely associated with GPT-4o's advanced logical capabilities. In the results of GPT-4, the lower RPTS scores are associated with erroneous reasoning and the model's failure to capture certain infomation. Conversely, the open-source models demonstrated a lack of logical robustness in their reasoning processes, leading to more pronounced decreases due to often generating irrelevant or illogical outputs. Despite these models' lower accuracy, their RPTS scores were not significantly impacted. We hypothesize that this is due to two primary reasons: 1. Disconnection between the inference outcomes and the intended targets. While the models initially could reason based on the specified targets, they gradually lost focus on the targets as the number of reasoning steps increased, resulting in conclusions that diverged from the intended data targets. 2. Recurrent generation of identical sentences. Across various sizes, the open-source models consistently produced repetitive reasoning that, while logically sound, failed to reach the desired conclusions. These factors led to reduced accuracy but did not substantially affect the logical integrity of the inferences, as reflected in the relatively high RPTS scores.

Step Analysis To further identify the causes of errors in our model, we initiated an analysis from the perspective of inference steps. We set $\alpha = 0$ and varied β at values of 1, 2, 3, and 4 to compute the average RPTS score. Figure 6 displays the relationship between RPTS scores and β across two languages. As evident from the figure, with the exception of GPT-4o, RPTS scores at $\beta = 1$ are unsatisfactory across all models. This indicates that the models encounter issues at the initial inference step, where conclusions are drawn directly from the visual and textual clues, leading to subsequent errors in reasoning. To further explore the specific causes, we calculated the average RPTS scores derived separately from visual and textual clues. The results, as shown in Table 3, reveal that open-source models still lack sufficient capabilities in image processing. They fail to derive necessary information from images for subsequent reasoning tasks based on specific inferential questions.

Capability Analysis Appendix B shows the accuracy of each model on six different abilities. The majority of models exhibit relatively uniform performance across six competencies in english.

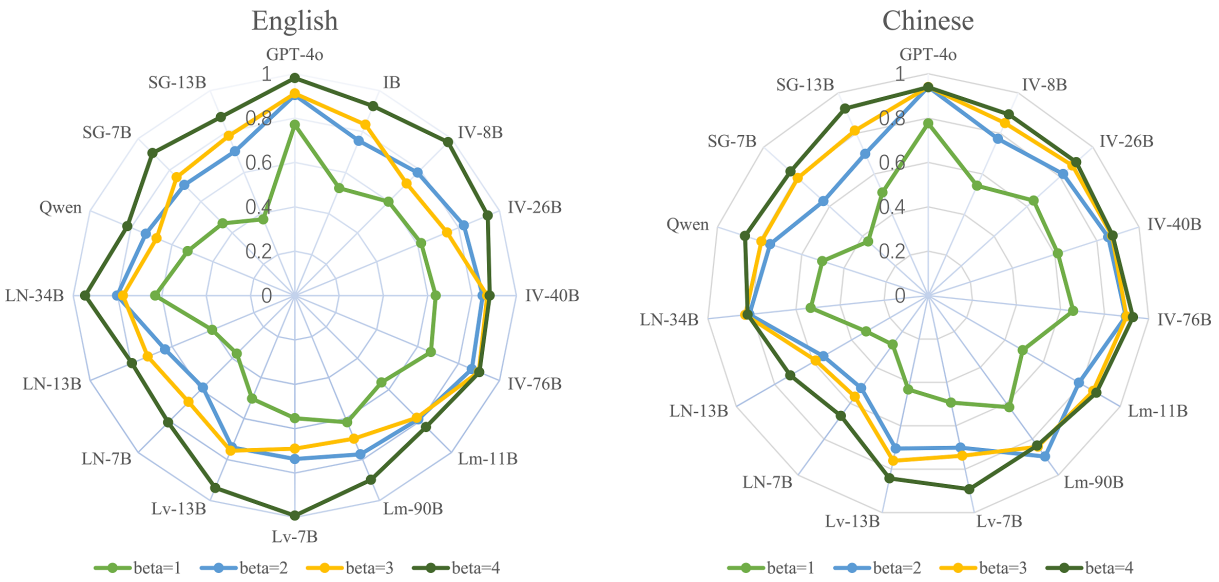

Figure 6: RPTS scores for $\beta \in \{1, 2, 3, 4\}$ and $\alpha = 0$. **IB**: InstructBLIP;**IV**: InternVL2;**Lv**: Llava-v1.5;**LN**: Llava-Next;**Qwen**: Qwen-VL-Chat;**SG**: ShareGPT4V;

Models	English		Chinese	
	Image	Text	Image	Text
IV-8B	0.50	0.76	0.54	0.94
Lv-7B	0.42	0.78	0.40	0.67
LN-7B	0.36	0.52	0.22	0.58
SG-7B	0.35	0.62	0.30	0.70
Lm-11B	0.52	0.83	0.4	1.0
IB	0.40	0.69	-	-
Lv-13B	0.41	0.66	0.37	0.78
LN-13B	0.36	0.57	0.29	0.73
Qwen	0.45	0.74	0.45	0.66
SG-13B	0.18	0.57	0.42	0.75
IV-26B	0.53	0.80	0.57	0.84
LN-34B	0.54	0.80	0.52	0.72
IV-40B	0.61	0.90	0.60	0.88
IV-76B	0.60	0.92	0.60	0.87
Lm-90B	0.58	0.79	0.6	0.75
GPT-4o	0.72	0.88	0.75	0.96

Table 3: RPTS score for drawing conclusions from visual clues or textual clues. **IB**: InstructBLIP; **IV**: InternVL2; **Lm**: Llama-3.2; **Lv**: Llava-v1.5; **LN**: Llava-Next; **Qwen**: Qwen-VL-Chat; **SG**: ShareGPT4V;

ShareGPT4V underperforms in OCR, which likely stems from a lack of targeted training data for these specific tasks. Surprisingly, open-source models, regardless of their parameter sizes, do not demonstrate particularly exceptional performance in the task of image comparison. Moreover, there is a significant gap between these models and GPT-4, which can be closely attributed to the fact that the training data used by most open-source models typically contains only a single image. Conversely, the performance of various open-source models markedly declines, displaying significant deficiencies in certain capabilities in Chinese. This is observed despite some models, such as Llava-Next-34B and Qwen-VL-Chat, leveraging LLM with robust capabilities in Chinese. This trend indicates that existing training methodologies fall short in translating a model's multimodal abilities from English into other languages.

5 Conclusion

In this paper, we introduce RPTS-Eval, a benchmark specifically designed to meticulously examine the reasoning processes of models. We also define three types of relationships between modalities in multimodal reasoning. Furthermore, we propose a new metric, RPTS, aimed at addressing issues where incorrect reasoning still results in correct outcomes, thereby facilitating a detailed analysis of model reasoning. Our results indicate that current open-source Large Visual Language Models struggle to derive necessary conclusions from images for subsequent reasoning. We also observed a significant disparity in the capabilities of models between Chinese and English contexts, suggesting that existing training methodologies fall short in transferring multimodal abilities from English to other languages.

6 Limitation

The main limitation of this paper lies in the scale of the data. Due to the lack of automated methods for constructing the multimodal reasoning

data presented in this paper, and the cost of manual construction is high, resulting in a relatively small dataset. Moreover, the relationships between modality in the RPTS-Eval data mainly correspond to related without interference, with insufficient data for the other two relationship types. Additionally, the proposed RPTS metric focuses solely on the dimension of logicality, and future work may explore evaluations across more dimensions.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems, 35:23716–23736.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023a. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. Preprint, arXiv:2308.12966.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023b. Touchstone: Evaluating vision-language models by language models. arXiv e-prints, pages arXiv–2308.

Yonatan Bitton, Hritik Bansal, Jack Hessel, Rulin Shao, Wanrong Zhu, Anas Awadalla, Josh Gardner, Rohan Taori, and Ludwig Schmidt. 2023. Visit-bench: A dynamic benchmark for evaluating instruction-following vision-and-language models. In Advances in Neural Information Processing Systems, volume 36, pages 26898–26922. Curran Associates, Inc.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. arXiv preprint arXiv:2311.12793.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821.

Cheng-Han Chiang and Hung-Yi Lee. 2023. Can large language models be an alternative to human evaluations? In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15607–15631.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in Neural Information Processing Systems, 36.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6556–6576.

Olga Golovneva, Moya Peng Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In The Eleventh International Conference on Learning Representations.

Xiaotian Han, Quanzeng You, Yongfei Liu, Wentao Chen, Huangjie Zheng, Khalil Mrini, Xudong Lin, Yiqi Wang, Bohan Zhai, Jianbo Yuan, Heng Wang, and Hongxia Yang. 2023. Infimm-eval: Complex open-ended reasoning evaluation for multi-modal large language models. Preprint, arXiv:2311.11567.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 13299–13308.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In International conference on machine learning, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 26296–26306.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. Advances in neural information processing systems, 36.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2511–2522.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In European Conference on Computer Vision, pages 216–233. Springer.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, pages 3195–3204.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. Preprint, arXiv:2303.08774.

Archiki Prasad, Swarnadeep Saha, Xiang Zhou, and Mohit Bansal. 2023. Receval: Evaluating reasoning chains via correctness and informativeness. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 10066–10086.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. Preprint, arXiv:2308.02490.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9556–9567.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6720–6731.

A Models' Detail

Model	Architecture		
	Vision	Language	Other
InstructBLIP	ViT-G	Vicuna-13B	Q-Former
InternVL2-8B	InternViT-300M	InternLM2-7B	MLP
InternVL2-26B	InternViT-6B	InternLM2-20B	MLP
InternVL2-40B	InternViT-6B	InternLM2-34B	MLP
InternVL2-76B	InternViT-6B	Llama3-70B	MLP
Llama-3.2-Vision-11B	Llama-Vision	Llama3.1-11B	Cross-Attention
Llama-3.2-Vision-90B	Llama-Vision	Llama3.1-90B	Cross-Attention
Llava-v1.5-7B	ViT-L	Vicuna-7B	MLP
Llava-v1.5-13B	ViT-L	Vicuna-13B	MLP
Llava-Next-7B	ViT-L	Vicuna-7B	MLP
Llava-Next-13B	ViT-L	Vicuna-13B	MLP
Llava-Next-34B	ViT-L	Yi-34B	MLP
Qwen-VL-Chat	ViT-bigG	Qwen-7B	-
ShareGPT4V-7B	ViT-L	Vicuna-7B	MLP
ShareGPT4V-13B	ViT-L	Vicuna-13B	MLP

Table 4: Open-source models' architecture

B Capability Accuracy

Models	English						Chinese					
	IC	Rec	OCR	SA	Com	Math	IC	Rec	OCR	SA	Com	Math
Lv-7B	0.40	0.49	0.47	0.52	0.53	0.36	0.20	0.25	0.19	0.26	0.24	0.22
LN-7B	0.00	0.34	0.34	0.27	0.34	0.29	0.00	0.07	0.06	0.06	0.09	0.03
Qwen	0.50	0.43	0.40	0.43	0.46	0.40	0.40	0.28	0.15	0.20	0.28	0.27
SG-7B	0.55	0.40	0.36	0.37	0.42	0.39	0.20	0.20	0.15	0.19	0.24	0.15
IV-8B	0.45	0.55	0.51	0.54	0.53	0.53	0.35	0.38	0.36	0.32	0.37	0.39
Lm-11B	0.5	0.58	0.51	0.58	0.58	0.47	0.2	0.29	0.36	0.21	0.29	0.25
IB	0.25	0.42	0.43	0.45	0.42	0.36	-	-	-	-	-	-
Lv-13B	0.30	0.43	0.39	0.42	0.49	0.40	0.35	0.31	0.25	0.38	0.35	0.22
LN-13B	0.05	0.37	0.33	0.40	0.36	0.27	0.05	0.14	0.10	0.11	0.14	0.12
SG-13B	0.30	0.34	0.29	0.35	0.33	0.27	0.45	0.26	0.29	0.25	0.26	0.29
IV-26B	0.40	0.57	0.63	0.53	0.57	0.63	0.10	0.48	0.54	0.45	0.48	0.43
LN-34B	0.50	0.62	0.57	0.56	0.59	0.54	0.20	0.37	0.40	0.33	0.38	0.35
IV-40B	0.50	0.69	0.69	0.65	0.69	0.71	0.45	0.55	0.51	0.42	0.47	0.54
IV-76B	0.55	0.72	0.75	0.65	0.67	0.77	0.35	0.58	0.54	0.53	0.54	0.61
Lm-90B	0.5	0.66	0.72	0.73	0.66	0.63	0.35	0.53	0.63	0.50	0.49	0.55
GPT-4o	0.95	0.85	0.89	0.80	0.83	0.84	0.95	0.85	0.89	0.80	0.83	0.84

Table 5: Accuracy for each capability filtered by RPTS score. For each column, the highest, the second, and the third highest figures are highlighted by green, orange and pink backgrounds. **IB**: InstructBLIP; **IV**: InternVL2; **Lm**: Llama-3.2; **Lv**: Llava-v1.5; **LN**: Llava-Next; **SG**: ShareGPT4V;

C Examples of Filtered-out Reasoning

Statement:
Person A has not violated the law.
Context:
The vehicle driven by person A is as shown in image 0.

Texture clues:
(0): Person A possesses a Category 1 driving license.
(1): Driving non-commercial vehicles with seven seats or less requires a Category 1 driver's license; otherwise, it is illegal.
(2): Driving vehicles with more than seven seats, excluding freight vehicles, requires a Category 2 driving license; otherwise, it is illegal.
(3): Operating a freight vehicle requires a Category 3 driving license; otherwise, it is illegal.

Model's reasoning:
Image0 -> Conclusion0: The vehicle driven by Person A has more than seven seats.
Text0 -> Conclusion1: A Category 1 driver's license is required to drive non-commercial vehicles with seven seats or less.
Conclusion0 + Conclusion1 -> Conclusion2: The vehicle driven by Person A requires a Category 2 driver's license.
Text1 -> Conclusion3: Operating a freight vehicle requires a Category 3 driver's license.
Conclusion0 + Conclusion3 -> Conclusion4: The vehicle driven by Person A is not a freight vehicle.
Conclusion2 + Conclusion4 -> Conclusion5: The vehicle driven by Person A does not require a Category 1, Category 2, or Category 3 driver's license.
Conclusion5 -> Conclusion6: Person A has not violated the law by driving the vehicle shown in the image.
Rpts Scores: 0.36

Statement:
The proposal was approved.
Context:
The company is hosting a meeting where a significant proposal was put to a vote.
Image 0 depicts the scene of the voting session.

Texture clues:
(0): The participants in the meeting constitute three-fifths of the entire company's workforce.
(1): The proposal can only pass if it receives support from more than 50% of the company's total employees.

Model's reasoning:
Image0 -> Conclusion0: According to image 0, we can infer that the proposal received support from eight participants in the meeting. Six people have raised their hands in favor, while two have not.
Conclusion0 + Text0 -> Conclusion1: These eight participants constitute three-fifths of the company's total workforce, as stated in Text0.
Text1 + Text2 -> Conclusion2: For the proposal to pass, more than 50% of the company's total employees must support it. Given that three-fifths of the workforce participated in the meeting, the number of total employees can be calculated. Let's assume the company's total workforce is represented by x. 3x/5 = 8 -> x = 40/3. Since 40/3 is not a whole number and the total workforce must be an integer, there may be rounding involved. If we approximate x as 14 (nearest whole number): Three-fifths of 14 employees is approximately 8.4 (closer to 8 considering participants). For the proposal to pass, it needs more than half of the entire company's employees to agree:14/2=7. Thus, more than 7 employees need to support the proposal.
Conclusion1 + Conclusion2 -> Conclusion3: 6 supporters (as shown in Image 0) out of 14 (approx total workforce) do not meet the requirement of more than 50% (i.e., more than 7 employees) for passing the proposal. Therefore, these textual and visual clues disagree with the given statement.
Rpts Scores: 0.43

Figure 7: Examples of filtered sample