RPTS: Tree-Structured Reasoning Process Scoring for Faithful Multimodal Evaluation

Haofeng Wang[1], **Yu Zhang**[1]*

[1]Harbin Institute of Technology, Harbin, China
hfwang@ir.hit.edu.cn, zhangyu@ir.hit.edu.cn

Abstract

Large Vision-Language Models (LVLMs) excel in multimodal reasoning and have shown impressive performance on 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 assess the reasoning process, their methods are often overly 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 intermodal relationships on reasoning. To address this issue, we propose the Reasoning Process Tree Score (RPTS), a tree structure-based metric to assess reasoning processes. Specifically, we organize the reasoning steps into a reasoning tree and leverage its hierarchical information to assign weighted faithfulness scores to each reasoning step. By dynamically adjusting these weights, RPTS not only evaluates the overall correctness of the reasoning, but also pinpoints where the model fails in the reasoning. To validate RPTS in real-world multimodal scenarios, we construct a new benchmark, RPTS-Eval, comprising 374 images and 390 reasoning instances. Each instance includes reliable visual-textual clues that serve as leaf nodes of the reasoning tree. Furthermore, we define three types of intermodal relationships to investigate how intermodal interactions influence the reasoning process. We evaluated representative LVLMs (e.g., GPT4o, 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 the advancement of research in the field of multimodal reasoning.

Code & Datasets — https://github.com/wang-hao-feng/RPTS

Extended version — https://arxiv.org/abs/2511.06899

Introduction

Recent advances in multimodal foundation models have demonstrated increasingly sophisticated capabilities in the combination of visual and textual information (OpenAI et al. 2024). However, as these models begin to assist in evidentiary reasoning tasks such as criminal case analysis - where establishing reliable connections between surveillance footage (visual modality), forensic reports (textual modality), and other evidence is crucial, and where conclusions must follow rigorous, verifiable reasoning chains - two critical questions emerge: 1. Can current evaluations distinguish between logically valid reasoning and coincidentally correct conclusions? 2. Do existing frameworks capture the non-linear, cross-modal reasoning required to resolve conflicting evidence?

Most existing benchmarks focus solely on task accuracy through multiple choice or short answer formats (Yu et al. 2024; Yue et al. 2024), completely ignoring the reasoning process. This approach fails to detect when models arrive at correct conclusions through flawed reasoning, a phenomenon we call "right answers for wrong reasons" , as illustrated on the left side of Figure 1. The few works that examine reasoning processes (Golovneva et al. 2023; Prasad et al. 2023) typically adopt an oversimplified linear evaluation framework. These approaches are fundamentally mismatched to the complex, non-linear nature of real-world reasoning, where multimodal evidence may appear conflicting yet collectively support valid conclusions. To address these limitations, we introduce a novel evaluation metric: Reasoning Process Tree Score (RPTS), designed to assess multimodal reasoning processes. The core innovation of RPTS lies in its tree-structured representation of reasoning, where leaf nodes correspond to atomic evidence units (visual or textual) and non-leaf nodes capture the hierarchical derivation of intermediate conclusions. This structure inherently accommodates the non-linear interactions characteristic of multimodal reasoning. Furthermore, RPTS incorporates two key hyperparameters whose adjustable values enable precise quantification of both global and local logical consistency, thereby facilitating accurate error localization within the reasoning chain.

To support comprehensive evaluation using RPTS, we developed the RPTS-Eval benchmark comprising 390 carefully constructed reasoning instances. Figure 2 shows examples of RPTS-Eval. Each instance contains complete and reliable multimodal atomic evidence for building reasoning trees. To systematically investigate how inter-modal relationships affect reasoning, we defined three distinct modality interaction types: guided (related without interference),

*Corresponding author

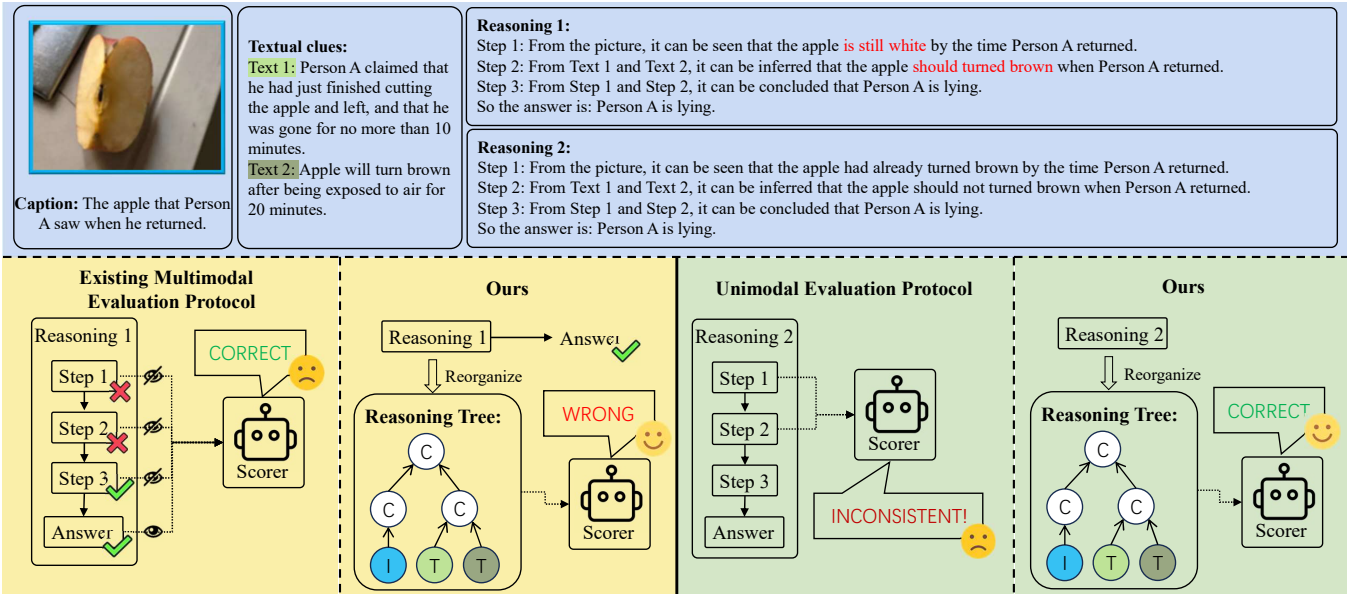

Figure 1: Comparison between Unimodal, existing Multimodal benchmarks and our RPTS. **Left**: Current multimodal benchmarks fail to detect instances where reasoning errors are present, yet the answer remains correct. **Right**: The unimodal approach is unable to handle reasoning involving conflicting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

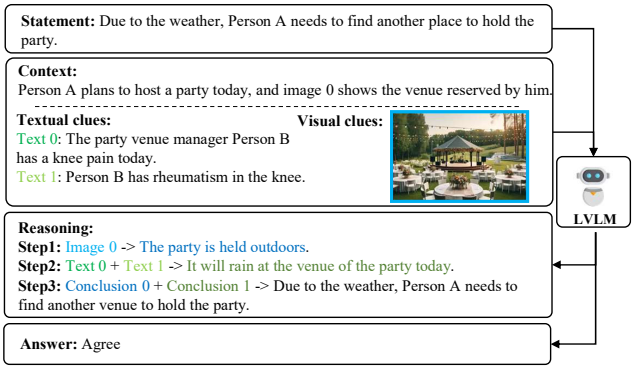

Figure 2: An example of RPTS-Eval.

adversarial (related with interference), and independent (unrelated), with each instance manually annotated accordingly. Experimental results on RPTS-Eval demonstrate RPTS's capability to identify flawed reasoning and precisely localize errors, while revealing significant limitations in current LVLMs' reasoning abilities. The primary contributions of our work can be summarized as follows:

- We introduce a new metric, RPTS, for detecting correct conclusions based on faulty reasoning and genuinely logical reasoning processes, reflecting both overall and local logic of reasoning, achieving error localization.
- We constructed RPTS-Eval, a novel benchmark for multimodal reasoning evaluation. Compared to existing datasets, RPTS-Eval provides reliable multimodal annotations of atomic evidence that facilitate a rigorous assessment of reasoning processes.
- We define three types of relationships between modalities in reasoning, which clarify the classification of multimodal reasoning.
- We conducted extensive experiments with our RPTS-Eval. The results reveal that current open-source LVLMs have difficulty drawing conclusions from images for further inference and show varying performance across different languages.

Related Work

MLLM Evaluation Benchmarks Classic multimodal benchmarks typically assess the specific reasoning abilities of the models. For example, 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 evaluate the comprehensive capabilities of a model, researchers have proposed various benchmarks, such as MMBench(Liu et al. 2025), SEED-Bench(Li et al. 2024), MM-Vet(Yu et al. 2024), 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 reasoning and its evaluation method cannot exclude cases where incorrect reasoning leads to a correct answer.

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.

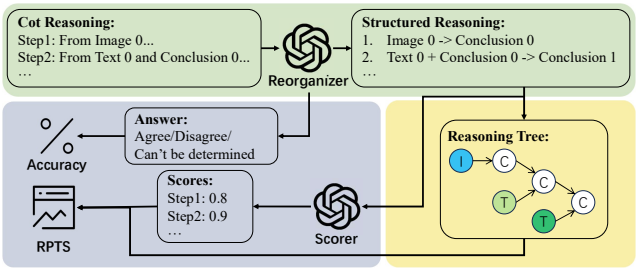

Figure 3: The calculation process of RPTS and accuracy.

RPTS

The computation of RPTS consists of two stages: Reasoning Parsing and Metric Calculation. Figure 3 illustrates the workflow of RPTS alongside the accuracy computation.

Reasoning Parsing

To construct the reasoning tree, we first parse the model's reasoning into a structured format: "[PREMISE] + [PREMISE] + ... → [CONCLUSION]", where '[PREMISE]' can be derived from visual clues, textual clues, or intermediate conclusions from prior steps. However, existing open-source MLLMs cannot strictly adhere to this output format. To address this, we first employ chain-of-thought (CoT) prompting to guide the model in generating step-by-step reasoning with explicit premises. Subsequently, we use GPT-4 to reformat the reasoning into a structured, easily parsable representation. The parsed reasoning results can also be utilized for accuracy computation.

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. 2024; Han et al. 2023) have demonstrated LLMs' effectiveness in assessing model reasoning. Therefore, we utilize a LLM 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 calculate the semantic similarity of conclusions directly derived from images against the ground truth. For other reasoning, we input the premises and conclusion into LLM 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.

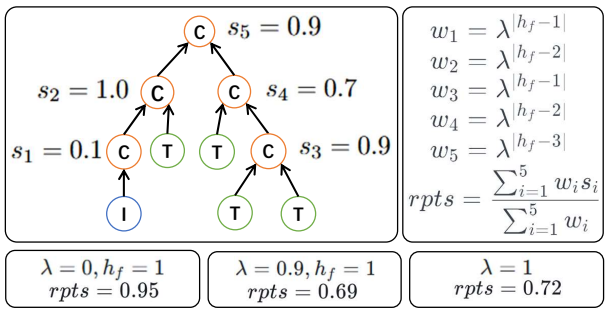

Figure 4: Examples of different hyperparameter settings for RPTS. C, I and T respectively represent conclusion, visual clue and textual clue.

Reasoning Process Tree Score

Considering the unique structure of reasoning, we can model the process as an reasoning tree, as depicted in Figure 4. 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 λ and h_f, is used to weight each inferential step. The weight assigned to n_i is defined as

$$w_i = \lambda^{|h_f - h|} \tag{1}$$

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. h_f is an integer that indicates the step most focused on by the RPTS, meaning that the weight is maximized at step h_f , with the weights decaying along the reasoning tree centered around this step. λ is the decay factor, which controls the speed at which the reasoning weight decays. 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 λ and h_f, we can finely tune the emphasis on global versus local aspects of the inference process.

Figure 4 illustrates the scoring outcomes under three different settings of λ and h_f. If we aim to assign equal importance to each reasoning step, we can set λ to 1, as shown

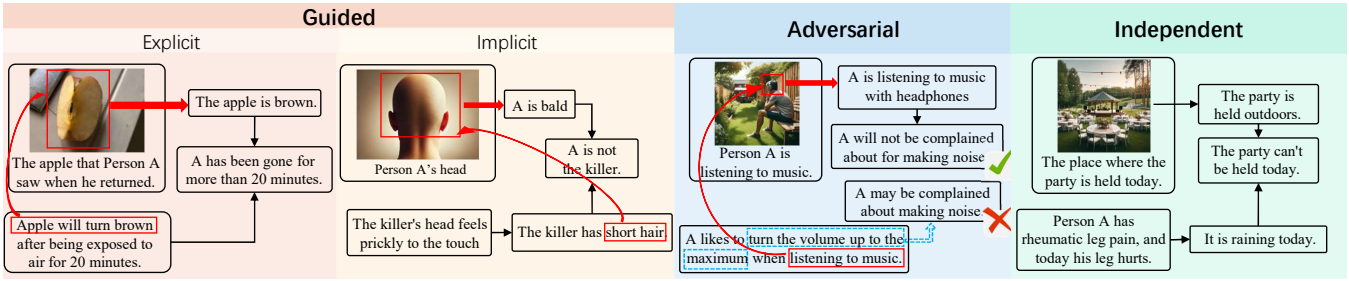

Figure 5: Three types of relationships between modalities. The filled arrow represents the relation between image and text, and the outlined arrow represents the interference between them.

in the top-right corner of Figure 4. In this case, RPTS represents the average score of all reasoning steps. Conversely, if we only wish to focus on the first step, i.e., reasoning at height 1, as depicted in the bottom-left corner, we set λ to 0 and h_f to 1. Under this configuration, only the score of the first step contributes to the RPTS computation. In the main experiment presented in Table 3, the hyperparameters we configured are consistent with those shown in the bottom-right corner of Figure 4. We focused more on the first step in reasoning, so we set $h_f = 1$. As shown in Table 2, most of the reasoning trees in RPTS-Eval have a height of no more than 7, and we want the weight of each reasoning step to be no less than 0.5 when calculating the RPTS. Based on Formula 1, we can calculate $\lambda > \sqrt[|h_f-7|]{0.5} \approx 0.891$. Therefore, we selected $\lambda = 0.9$.

RPTS-Eval

To enable a fine-grained analysis of models' reasoning capabilities, we propose RPTS-Eval, a novel multimodal reasoning evaluation benchmark. Each reasoning instance in RPTS-Eval contains reliable visual-textual clues that serve as leaf nodes of the reasoning tree, facilitating structured reasoning assessment.

Data Collection

We aim to developing a high-quality multimodal reasoning evaluation benchmark, using a meticulously designed methodology to assess model reasoning performance. Each sample in RPTS-Eval can 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 Inspiration. 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. 2024), we then require GPT-4 to generate reasoning stories encompassing six types of capabilities, based on the given examples. However, we define two distinct capabilities that differ from MM-Vet: Image Comparison (IC) and Spatial Awareness (SA). The remaining four capabilities—Recognition (Rec), OCR, Commonsense Reasoning (Com), and Math—are identical to those defined in MM-Vet. The specific definitions of IC and SA are as follows:

- **Image Comparison(IC):** The model compares two images to spot similarities or differences. This is a basic human skill, as we learn a lot about the world by observing and comparing things.
- **Spatial Awareness(SA):** This covers spatial skills, like recognizing fixed positions or understanding how objects relate to each other from different viewpoints.

As mentioned above, the stories generated by GPT-4 lack logical consistency. Therefore, annotators only draw inspiration from these stories rather than using them directly, thereby reducing the difficulty of story design. For example, GPT-4 generates a story set on a rainy day, and our annotators draw inspiration from this, such as the idea of 'rain,' and design reasoning tasks based on that inspiration. For instance, they might infer whether the box is open or closed based on whether there's water inside, or they might reason about whether outdoor activities can continue based on certain pre-rain features.

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:** Check how statements, context, visuals, and reasoning connect, and score them to ensure

Benchmark	Size	Images	Answer Format	Metric	Evaluate Reasoning
MMMU(Yue et al. 2024)	11.5K	12.5K	Option/Open Answer	Accuracy	✗
MM-Vet(Yu et al. 2024)	218	200	Shot answer	GPT4-score	✗
InifMM-Eval(Han et al. 2023)	279	342	Reasoning	GPT4-score	✓
RPTS-Eval(Ours)	390	374	Reasoning	Accuracy+RPTS	✓

Table 1: The comparison between RPTS-Eval and other existing benchmarks.

strong logic.

- **Multimodality:** Remove samples that don't require both visual and text clues for reasoning (single-modality solvable).
- **Subjectivity and Discrepancy Check:** Discard or edit overly subjective data or cases where validator reasoning clashes with ground truth.
- **Missing or Redundant abilities:** Validators flag missing or unnecessary annotated reasoning abilities.

We excluded data where there was disagreement between the two validators as well as those with low logical scores.

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 5.

- **Guided**: 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.
- **Adversarial**: In some cases, one modality can negatively influence information extraction from another, either by leading to irrelevant/incorrect data or by preventing useful information from being retrieved at all.
- **Independent**: Modalities don't influence each other, information must be gathered separately from each for reasoning.

Dataset Statistics

In summary, our RPTS-Eval benchmark comprises 390 inferences linked to a total of 374 images. Table 2 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 guided, 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. For a comparison with other benchmarks, please refer to Table 1.

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			
Guided	84.62%	Adversarial	6.92%
Independent	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 2: Key statistics of the RPTS-Eval benchmark. As each reasoning instance need one or more capabilities, the sum of percentage is larger than 100%.

Experiments

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), InternVL2(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.

Scorer Selection

To select an appropriate scoring model, we randomly sampled 200 reasoning instances and manually scored them. Concurrently, we selected five distinct models of varying

Models	English			Chinese		
	Acc	**RPTS**↑	**Acc**$_{filtered}$	**Acc**	**RPTS**↑	**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 3: Results of different models on RPTS-Eval with cot prompt. We set $\lambda = 0.9$, $h_f = 1$ when calculate RPTS. For each column, the highest, the second, and the third highest figures are highlighted by **bold**, underline and star*. **Acc**: Accuracy.

Model	$\overline{\Delta}$
Qwen2-7B	0.216
Llama-3-8B	0.231
Qwen2-72B	0.143
Llama-3-70B	0.152
GPT-4	**0.095**

Table 4: Mean absolute error ($\overline{\Delta}$) between different LLM scores and human scores

types and sizes as potential scorers, evaluating their performance against human-assigned scores. Table 4 presents the **M**ean **A**bsolute **E**rror (MSE) between the scores generated by these models and those assigned by humans. Based on the minimal discrepancy observed, we opted for GPT-4 as our designated scoring model.

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. When reasoning, we set the temperature of each model to 0 and use greedy decoding.

Results and Analysis

Table 3 presents model performance on RPTS-Eval. Beyond inference accuracy and mean RPTS scores, we applied an RPTS-based filter (score ¡ 0.5) to exclude cases where correct conclusions arose from flawed reasoning. Results show all models experienced accuracy declines, with GPT-4o least affected—consistent with its stronger logical capacity. 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. In addition, Appendix B shows the performance of the model on six capabilities.

Step Analysis. To further identify the causes of errors in our model, we initiated an analysis from the perspective of inference steps. We set $\lambda = 0$ and varied h_f at values of 1, 2, 3, and 4 to compute the average RPTS score. Figure 6 displays the relationship between RPTS scores and h_f across two languages. As evident from the Figure 6, with the exception of GPT-4o, RPTS scores at $h_f = 1$ are unsatisfactory across all models. This indicates that the models encounter issues at the initial inference step, where conclu-

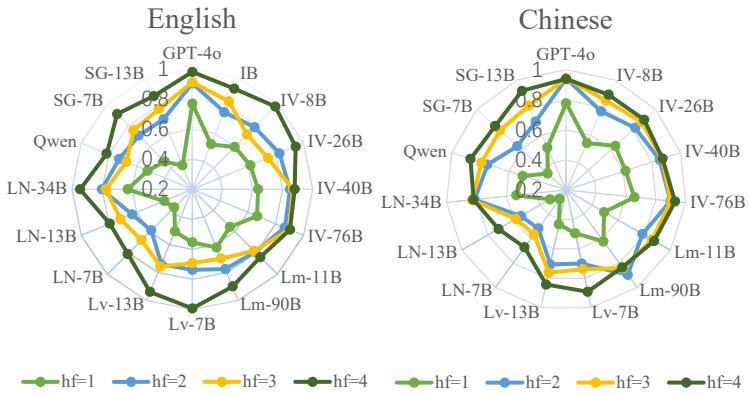

Figure 6: RPTS scores for $h_f \in \{1, 2, 3, 4\}$ and $\lambda = 0$. **IB**: InstructBLIP; **IV**: InternVL2; **Lv**: Llava-v1.5; **LN**: Llava-Next; **Qwen**: Qwen-VL-Chat; **SG**: ShareGPT4V;

Models	English		Chinese	
	V	T	V	T
InternVL2-8B	0.50	0.76	0.54	0.94
Llava-1.5-7B	0.42	0.78	0.40	0.67
Llava-Next-7B	0.36	0.52	0.22	0.58
ShareGPT4V-7B	0.35	0.62	0.30	0.70
Llama-3.2-11B	0.52	0.83	0.4	1.0
InstructBLIP	0.40	0.69	-	-
Llava-1.5-13B	0.41	0.66	0.37	0.78
Llava-Next-13B	0.36	0.57	0.29	0.73
Qwen-VL-Chat	0.45	0.74	0.45	0.66
ShareGPT4V-13B	0.18	0.57	0.42	0.75
InternVL2-26B	0.53	0.80	0.57	0.84
Llava-Next-34B	0.54	0.80	0.52	0.72
InternVL2-40B	0.61	0.90	0.60	0.88
InternVL2-76B	0.60	0.92	0.60	0.87
Llama-3.2-90B	0.58	0.79	0.6	0.75
GPT-4o	0.72	0.88	0.75	0.96

Table 5: RPTS score for drawing conclusions from visual clues(**V**) or textual clues(**T**).

sions 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 5, 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.

Sensitivity Analysis To further investigate the impact of different λ and h_f values on RPTS and the correctness of reasoning, we conducted a sensitivity analysis using the reasoning results from InternVL-26B. Table 6 presents the RPTS values and the proportion of filtered reasoning paths for various λ and h_f settings. From the table, it can be observed that small variations in λ do not significantly alter the RPTS values or the proportion of filtered reasoning paths.

h_f \ λ	0.2	0.4	0.6	0.8	1.0
1	0.647	0.671	0.690	0.703	0.713
2	0.768	0.743	0.728	0.719	0.713
3	0.733	0.733	0.727	0.720	0.713
4	0.733	0.734	0.729	0.721	0.713

h_f \ λ	0.2	0.4	0.6	0.8	1.0
1	18.21	16.15	14.36	10.51	10.26
2	8.97	9.74	9.23	8.97	10.26
3	10.00	9.74	10.51	9.49	10.26
4	10.26	9.49	9.49	9.23	10.26

Table 6: Sensitivity analysis of RPTS: values (top) and percentage of low-score correct answers (bottom) under various λ and h_f.

However, changes in h_f lead to notable differences, particularly when λ is small. This aligns with our design intention: a smaller λ reduces the influence of non-h_f steps, thereby making RPTS more closely reflect the score of the h_f step.

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.

Acknowledgments

We sincerely thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Natural Science Foundation of China (No. 62476066).

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023a. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv:2308.12966.

Bai, S.; Yang, S.; Bai, J.; Wang, P.; Zhang, X.; Lin, J.; Wang, X.; Zhou, C.; and Zhou, J. 2023b. TouchStone: Evaluating Vision-Language Models by Language Models. *arXiv e-prints*, arXiv–2308.

Bitton, Y.; Bansal, H.; Hessel, J.; Shao, R.; Zhu, W.; Awadalla, A.; Gardner, J.; Taori, R.; and Schmidt, L. 2023. VisIT-Bench: A Dynamic Benchmark for Evaluating Instruction-Following Vision-and-Language Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 26898–26922. Curran Associates, Inc.

Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; 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*.

Chiang, C.-H.; and Lee, H.-Y. 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)*, 15607–15631.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2024. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems*, 36.

Fu, J.; Ng, S. K.; Jiang, Z.; and Liu, P. 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)*, 6556–6576.

Golovneva, O.; Chen, M. P.; Poff, S.; Corredor, M.; Zettlemoyer, L.; Fazel-Zarandi, M.; and Celikyilmaz, A. 2023. ROSCOE: A Suite of Metrics for Scoring Step-by-Step Reasoning. In *The Eleventh International Conference on Learning Representations*.

Han, X.; You, Q.; Liu, Y.; Chen, W.; Zheng, H.; Mrini, K.; Lin, X.; Wang, Y.; Zhai, B.; Yuan, J.; Wang, H.; and Yang, H. 2023. InfiMM-Eval: Complex Open-Ended Reasoning Evaluation For Multi-Modal Large Language Models. arXiv:2311.11567.

Li, B.; Ge, Y.; Ge, Y.; Wang, G.; Wang, R.; Zhang, R.; and Shan, Y. 2024. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13299–13308.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 216–233. Springer.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 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*, 2511–2522.

Marino, K.; Rastegari, M.; Farhadi, A.; and Mottaghi, R. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, 3195–3204.

OpenAI; Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; Avila, R.; Babuschkin, I.; Balaji, S.; Balcom, V.; Baltescu, P.; Bao, H.; Bavarian, M.; Belgum, J.; Bello, I.; Berdine, J.; Bernadett-Shapiro, G.; Berner, C.; Bogdonoff, L.; Boiko, O.; Boyd, M.; Brakman, A.-L.; Brockman, G.; Brooks, T.; Brundage, M.; Button, K.; Cai, T.; Campbell, R.; Cann, A.; Carey, B.; Carlson, C.; Carmichael, R.; Chan, B.; Chang, C.; Chantzis, F.; Chen, D.; Chen, S.; Chen, R.; Chen, J.; Chen, M.; Chess, B.; Cho, C.; Chu, C.; Chung, H. W.; Cummings, D.; Currier, J.; Dai, Y.; Decareaux, C.; Degry, T.; Deutsch, N.; Deville, D.; Dhar, A.; Dohan, D.; Dowling, S.; Dunning, S.; Ecoffet, A.; Eleti, A.; Eloundou, T.; Farhi, D.; Fedus, L.; Felix, N.; Fishman, S. P.; Forte, J.; Fulford, I.; Gao, L.; Georges, E.; Gibson, C.; Goel, V.; Gogineni, T.; Goh, G.; Gontijo-Lopes, R.; Gordon, J.; Grafstein, M.; Gray, S.; Greene, R.; Gross, J.; Gu, S. S.; Guo, Y.; Hallacy, C.; Han, J.; Harris, J.; He, Y.; Heaton, M.; Heidecke, J.; Hesse, C.; Hickey, A.; Hickey, W.; Hoeschele, P.; Houghton, B.; Hsu, K.; Hu, S.; Hu, X.; Huizinga, J.; Jain, S.; Jain, S.; Jang, J.; Jiang, A.; Jiang, R.; Jin, H.; Jin, D.; Jomoto, S.; Jonn, B.; Jun, H.; Kaftan, T.; Łukasz Kaiser; Kamali, A.; Kanitscheider, I.; Keskar, N. S.; Khan, T.; Kilpatrick, L.; Kim, J. W.; Kim, C.; Kim, Y.; Kirchner, J. H.; Kiros, J.; Knight, M.; Kokotajlo, D.; Łukasz Kondraciuk; Kondrich, A.; Konstantinidis, A.; Kosic, K.; Krueger, G.; Kuo, V.; Lampe, M.; Lan, I.; Lee, T.; Leike, J.; Leung, J.; Levy, D.; Li, C. M.; Lim, R.; Lin, M.; Lin, S.; Litwin, M.; Lopez, T.; Lowe, R.; Lue, P.; Makanju, A.; Malfacini, K.; Manning, S.; Markov, T.; Markovski, Y.; Martin, B.; Mayer, K.; Mayne, A.; McGrew, B.; McKinney, S. M.; McLeavey, C.; McMillan, P.; McNeil, J.; Medina, D.; Mehta,

A.; Menick, J.; Metz, L.; Mishchenko, A.; Mishkin, P.; Monaco, V.; Morikawa, E.; Mossing, D.; Mu, T.; Murati, M.; Murk, O.; Mély, D.; Nair, A.; Nakano, R.; Nayak, R.; Neelakantan, A.; Ngo, R.; Noh, H.; Ouyang, L.; O'Keefe, C.; Pachocki, J.; Paino, A.; Palermo, J.; Pantuliano, A.; Parascandolo, G.; Parish, J.; Parparita, E.; Passos, A.; Pavlov, M.; Peng, A.; Perelman, A.; de Avila Belbute Peres, F.; Petrov, M.; de Oliveira Pinto, H. P.; Michael; Pokorny; Pokrass, M.; Pong, V. H.; Powell, T.; Power, A.; Power, B.; Proehl, E.; Puri, R.; Radford, A.; Rae, J.; Ramesh, A.; Raymond, C.; Real, F.; Rimbach, K.; Ross, C.; Rotsted, B.; Roussez, H.; Ryder, N.; Saltarelli, M.; Sanders, T.; Santurkar, S.; Sastry, G.; Schmidt, H.; Schnurr, D.; Schulman, J.; Selsam, D.; Sheppard, K.; Sherbakov, T.; Shieh, J.; Shoker, S.; Shyam, P.; Sidor, S.; Sigler, E.; Simens, M.; Sitkin, J.; Slama, K.; Sohl, I.; Sokolowsky, B.; Song, Y.; Staudacher, N.; Such, F. P.; Summers, N.; Sutskever, I.; Tang, J.; Tezak, N.; Thompson, M. B.; Tillet, P.; Tootoonchian, A.; Tseng, E.; Tuggle, P.; Turley, N.; Tworek, J.; Uribe, J. F. C.; Vallone, A.; Vijayvergiya, A.; Voss, C.; Wainwright, C.; Wang, J. J.; Wang, A.; Wang, B.; Ward, J.; Wei, J.; Weinmann, C.; Welihinda, A.; Welinder, P.; Weng, J.; Weng, L.; Wiethoff, M.; Willner, D.; Winter, C.; Wolrich, S.; Wong, H.; Workman, L.; Wu, S.; Wu, J.; Wu, M.; Xiao, K.; Xu, T.; Yoo, S.; Yu, K.; Yuan, Q.; Zaremba, W.; Zellers, R.; Zhang, C.; Zhang, M.; Zhao, S.; Zheng, T.; Zhuang, J.; Zhuk, W.; and Zoph, B. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Prasad, A.; Saha, S.; Zhou, X.; and Bansal, M. 2023. ReCEval: Evaluating Reasoning Chains via Correctness and Informativeness. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10066–10086.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2024. MM-Vet: Evaluating Large Multimodal Models for Integrated Capabilities. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 57730–57754. PMLR.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; 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*, 9556–9567.

Zellers, R.; Bisk, Y.; Farhadi, A.; and Choi, Y. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6720–6731.