NPHardEval4V: Dynamic Evaluation of Large Vision-Language Models with Disentangled Effects of Vision Input

Anonymous ACL submission

Abstract

The reasoning capabilities of Large Vision-Language Models (LVLMs), also known as Multimodal Foundation Models, are a key area of research. We introduce **NPHardEval4V**, a dynamic benchmark designed to evaluate LVLMs' pure reasoning abilities, disentangling factors such as image recognition and instruction following. By converting textual descriptions of 900 reasoning tasks into image representations, this benchmark offers a focused assessment of reasoning. Our results highlight significant differences in performance across models, with LVLMs generally underperforming compared to Large Language Models (LLMs). We also analyze the effects of different prompting styles—visual, text, and combined visual-text—on reasoning performance, demonstrating the distinct influence of visual inputs. Unlike static benchmarks, NPHardEval4V employs a continuous update mechanism to mitigate overfitting and ensure more accurate evaluations over time. This benchmark provides a comprehensive view of LVLMs' reasoning abilities and is poised to guide future development in multimodal reasoning tasks. The dataset and code are available at https://anonymous.4open.science/r/NPHardEval4V-AB8E.

1 Introduction

The rise of Large Vision-Language Models (LVLMs) is a key step toward artificial general intelligence (AGI), building on the success of Large Language Models (LLMs) (Fan et al., 2023b; Zhao et al., 2023). LVLMs enhance multimedia interactions and cross-modal decision-making tools (Yin et al., 2023; Wang et al., 2024). For example, they can analyze visual representations of strings to determine the fewest operations needed to transform one into another, or find the shortest route on a map by interpreting the visual data. This reasoning ability allows LVLMs to make logical decisions based on available information.

Numerous benchmarks have been developed to assess the capabilities of LVLMs in various domains, such as visual question answering (Antol et al., 2015; Goyal et al., 2017; Marino et al., 2019; Hudson and Manning, 2019), Optical Character Recognition (Liu et al., 2023c), robustness (Zhao et al., 2024), hallucination (Liu et al., 2023a), and holistic overall performance (Xu et al., 2023; Li et al., 2023b; Liu et al., 2023b; Yue et al., 2023). However, despite the breadth of these evaluations, none of them specifically focus on assessing the pure reasoning abilities of LVLMs. Thereby, our benchmark aims to separate the effects of other factors, such as image recognition and instruction following, from the overall performance, allowing for precise reasoning evaluation. Furthermore, many of these benchmarks are static in nature, making them prone to overfitting and limiting their effectiveness in providing a comprehensive view of LVLMs' abilities (Chang et al., 2023). To address these limitations, there is a need for a dynamic benchmark that focuses on LVLMs' reasoning abilities and can update regularly to prevent overfitting.

The proposed **NPHardEval4V** benchmark is a visual counterpart of the NPHardEval benchmark, as presented in (Fan et al., 2023a). It comprises nine types of algorithmic problems, categorized into three polynomial time problems, three NP-complete problems, and three NP-hard problems. Each algorithmic problem consists of 100 instances with varying difficulty levels. To enable a direct comparison between LVLMs and LLMs, we have retained the problems from NPHardEval and converted their textual descriptions into visual representations.

Based on this benchmark, we investigate several research questions:

1. **Reasoning Performance Evaluation of LVLMs:** We first seek to evaluate the variation in reasoning abilities among different LVLMs and identify the factors that contribute to any

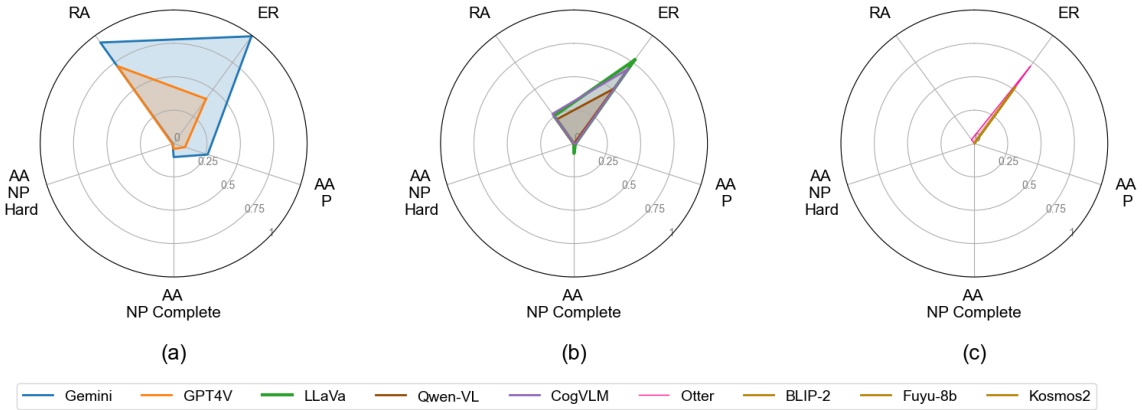

Figure 1: Large Vision-Language Models's performance on recognition (RA), Instruction-following (ER), and reasoning (AA) on polynomial time, NP-complete, and NP-hard problems.

observed performance gaps. By analyzing and separating the influence of recognition and instruction following, we aim to assess the core reasoning abilities of LVLMs and how they vary with the nature of complexity and difficulty of problems. This investigation will provide insights into the relative strengths and weaknesses of LVLMs in reasoning tasks. We provide the model details of the LVLMs we evaluate in Table 1.

2. **Effects of Vision Input on LVLMs' Performance:** We investigate how the inclusion of vision prompts influences LVLMs' reasoning abilities compared to pure text inputs. Through ablation studies of three prompt types, including "Figures with Limited Instructional Text Setup" (figure+limited_text), "Text-only Setup" (full_text_only), and "Vision-rich-text Setup" (figure+full_text), we aim to assess the impact of vision inputs on models' reasoning performance and examine how the mixed vision and text inputs can affect LVLMs' performance.

3. **Comparison of Reasoning Ability between LVLMs and LLMs** We investigate the difference between the current general development of LVLMs and LLMs in terms of reasoning abilities, for which we evaluate both types of models on the exact same set of questions. Specifically, we use visual representations and their textual counterparts as prompts to assess the models' performance in handling reasoning tasks. This comparison will provide valuable insights into the relative strengths and weaknesses of LVLMs and LLMs in reasoning tasks, and highlight areas for further research and development to improve the reasoning abilities of LVLMs.

2 Related Work

2.1 Large Vision-Language Models (LVLMs) and Their Reasoning Abilities

LVLMs have potentials in surpassing singularly moded LLMs and unlock new avenues for performing real-world applications (Wang et al., 2024; Wu et al., 2023). As summarized by (Yin et al., 2023), LVLMs are more intelligent, user-centric, and holistic as compared to their LLM counterparts.

Reasoning is one of the fundamental intelligent behaviors, essential for solving complex real-world tasks (Yu et al., 2023; Huang and Chang, 2022; Chen et al., 2024). However, even in text-centric settings, LLMs lack proper reasoning abilities, such as dealing with NP-hard (nondeterministic polynomial time) mathematical problems (Fan et al., 2023a). The quest to explore and improve reasoning capabilities of Strong AI particularly within LVLMs remains a challenge (Chen et al., 2023; Liu et al., 2023d; Gong et al., 2023).

In the domain of LVLMs, researchers have explored a range of techniques, such as instruction-tuning and prompt engineering, to enhance multimodal reasoning (Wang et al., 2024). The practice of instruction tuning, vital for in-context learning (ICL), has emerged as one of the key techniques in enhancing the reasoning abilities of these models (Tsimpoukelli et al., 2021; Alayrac et al., 2022; Koh et al., 2023; Guo et al., 2023).

2.2 Benchmarks of Large Vision-Language Models (LVLMs)

As the reasoning ability of LVLMs continues to advance, benchmarks have become instrumental in evaluating their performance and identifying areas that require improvement. Wang et al. (2024) suggest that a robust multimodal reasoning benchmark

must fulfill three key criteria: (1) the integration of multimodal information, (2) the categorization of reasoning, and (3) in-depth annotations of the reasoning steps (Wang et al., 2024).

Previous research has established numerous benchmarks to gauge the performance of LVLMs (Fu et al., 2023a,b; Li et al., 2023b; Liu et al., 2023a; Cao et al., 2023). However, they do not provide a clear picture of the pure reasoning ability of LVLMs because factors such as recognition, knowledge amount, instruction following, and others are all combined in the presented performances in these benchmarks. To address this research gap, our study proposes a dynamic assessment framework for evaluating the reasoning performance of LVLMs with a focus on disentangling the influence of visual input. In addition, we utilize the computational complexity hierarchy to rigorously assess the extent to which these LVLMs can achieve in reasoning tasks (Fan et al., 2023a).

3 Benchmark Construction

NPHardEval4V is built upon the NPHardEval Benchmark (Fan et al., 2023a), a benchmark that segments testing problems by computational complexity hierarchy into three complexity classes: P (polynomial time), NP-complete (nondeterministic polynomial-time complete), and NP-hard. These classifications serve to delineate the intrinsic difficulty and the computational resources required for solving tasks within each class, showcasing an increasing order of complexity. Table 2 presents the tasks included in the benchmark with the computational complexity class they belong to. Each task is further divided into 10 progressive difficulty levels with each difficulty level contains 10 instances. This granular difficulty grading enables a nuanced assessment of model performance across a spectrum of computational challenges, thereby offering valuable insights into the capabilities and limitations of current models' reasoning abilities.

NPHardEval4V also adopts the same three computational complexity hierarchy, transforming textual description of the problems in NPHardEval into visual representations[1]. This section outlines the transformation of the NPHardEval benchmark to suit the evaluation of LVLMs within NPHardEval4V, emphasizing the multimodal aspects of the tasks and the dynamic nature of the challenges.

[1]Robustness experiment on the benchmark is performed in (Fan et al., 2023a).

To enable the direct comparison with the text-only prompt input in the NPHardEval benchmark, we transform the text-only questions and represent the data part with figures. We further illustrate these transformations, with the data of the problem presented both textually and visually.

Graph Data Transformation The general construction of graph data problems involves providing LVLMs with both a textual description and a visual representation. For example, in the GCP (see Figure 2), LVLMs are given a textual prompt alongside a figure depicting the specific question in a graph, challenging them to reason across modalities. The figure is generated using Python based on the textual description of GCP: it includes the nodes and the edges between them. The figures in algorithmic questions such as SPP, GCP-D, TSP, and TSP-D are transformed from text similarly: these figures include graphs containing nodes and edges; in the case of TSP and TSP-D, the weights on the edges are also included.

```
Graph coloring refers to the problem of
    coloring vertices of a graph in such
     a way that no two adjacent vertices
     have the same color.
There are 7 vertices 1 to 7 in a graph.
    You may use 3 colors with alphabets
    from A, B, C,... to color the graph.
     The graph is provided.
```

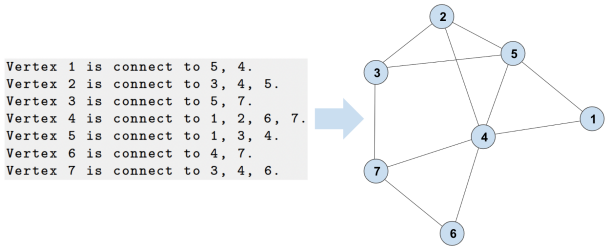

Figure 2: Image generation for GCP Example

Linear Data Transformation Similarly, linear data problems like the KSP (see Figure 3) are presented with both a prompt and a visual representation of items with associated weights and values. Visual representation of the KSP problem contains blocks with sizes corresponding to the weights of the items as described in the textual problem description. Specifically, each item is assigned a unique identifier (id) and a corresponding weight, which is used to determine the size of the block representing that item. The LVLMs are tasked with maximizing value within the constraints provided.

In the MSP problems, we generate figures that visually represent the availability of each person in

a calendar format corresponding to the textual description of the problem. Specifically, we create a calendar view that displays the availability of each person based on the input data. For algorithmic problems such as EDP and SAS, we do not employ a specialized visualization but simply display the relevant string of characters or string of numbers in the image.

```
The Knapsack Problem (KSP) asks whether
    a subset of items, each with a given
     weight and value, can be chosen to
    fit into a knapsack of fixed
    capacity, maximizing the total value
     without exceeding the capacity.
    Determine if a subset of items can
    be selected to fit into a knapsack
    with a capacity of 40, maximizing
    weight without exceeding the
    capacity. Item weights are presented
     by the top number after 'W:'. Item
    IDs are shown as numbers below the
    weights after 'id:'.
```

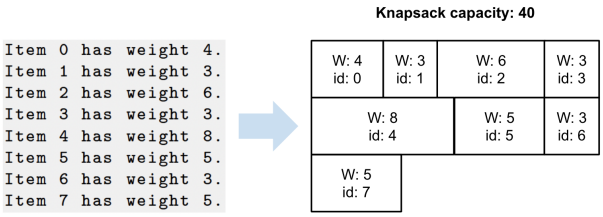

Figure 3: Image generation for KSP Example

4 Experimental Setting

To address the outlined research questions, our experimental setting systematically assesses the performance of various LVLMs in pure reasoning and investigates the impact of visual and textual prompts on LVLMs' performance. The LVLMs we test on are presented in Table 1.

4.1 Recognition and Reasoning Performance Evaluation

In the initial experiment, the objective is to assess the performance of LVLMs separating the effect on recognition and reasoning. Zero-shot prompts are utilized to evaluate the inherent recognition and reasoning capabilities of each model.

Recognition Experiment In order to precisely assess the capacity for reasoning in LVLMs, it is essential to first comprehend their proficiency in image recognition, as accurate reasoning can only be established upon correct recognition. To evaluate this, we prompt each LVLMs with a visual representation of the question and a textual representation of the same question for the model to determine whether they correspond to the same question. To mitigate the influence of randomness, each model is prompted five times with five different random seeds. The mean and variance of the model's responses are then calculated, and it is concluded that the model demonstrates recognition capabilities if it correctly identifies the correspondence between the visual and textual representations, i.e., answering "yes", for more than three times.

Reasoning Experiment 1, the Default Setup (Vision with Instructional Text) This experiment utilizes the standard setting of evaluating on the benchmark: a limited textual prompt including a general introduction to the task and the format in which the answer should be given; an image is presented that pertains to the specific question being asked. The models are then tasked to generate accurate responses. The performance of the LVLMs is evaluated based on their ability to correctly answer the questions, taking into account both their recognition and reasoning capabilities.

To accurately understand the pure reasoning ability, it is necessary to isolate the specific aspect of reasoning ability from other factors that may influence the overall performance: we employ a filtering process that removes datapoints where the model fails to recognize the given input or fails to provide a parsable result, which indicates a failure in instruction-following. After filtering out these two factors, we can obtain a more accurate picture of the LVLMs' pure reasoning ability.

4.2 Ablation on Multimodal prompts

The objective of this experiment is to explore the effect of prompt modality on the reasoning abilities of LVLMs. Humans usually use visual aids in solving complex problems. For example, when dealing with a Graph Coloring Problem (GCP), sketching a diagram of the graph and attempting different coloring schemes can aid in determining whether a valid solution exists. Therefore, in this experiment, we aim to evaluate whether the inclusion of visual representation in addition to the textual description of the specific question can also enhance the models' performance.

Reasoning Experiment 2, the Text-only Setup (Instructional and Data Text) To investigate the impact of visual representation on the reasoning

performance of LVLMs, it is necessary to first establish a baseline performance using pure textual prompts. In this experiment, we provide LVLMs with a pure textual description of the problem, including the description of a specific question. This will allow us to evaluate the models' reasoning abilities using only textual information, without any visual processing.

Reasoning Experiment 3, the Vision-rich-text Setup (Vision with Instructional and Data Text) In this experiment, we aim to investigate whether providing LVLM with both textual and visual representations of a problem can improve their performance in solving it. We present the LVLMs with a complete textual description of the problem, followed by an accompanying image that visually depicts the problem. Our hypothesis is that the image can serve as a supplementary source of information to the textual description, potentially enhancing the model's understanding of the problem and enabling it to generate more accurate solutions. By comparing the results of this experiment with those of the previous experiment where only textual descriptions were provided, we can assess the impact of visual representations on the reasoning abilities of LVLMs. This investigation will provide insights into how the combination of textual and visual inputs can affect LVLMs' performance in solving complex problems.

4.3 Comparison between LVLMs and LLMs on Reasoning

The objective of this experiment is to evaluate and compare the pure reasoning abilities of current state-of-the-art LVLMs and LLMs. To achieve this, we leverage the results from Reasoning Experiment 1, where LVLMs were evaluated on a set of reasoning problems with visual prompts, and compare them with the performance of LLMs on the exact same set of questions and problems, but with textual prompts instead of the visual ones.

4.4 Evaluation Metrics

Our evaluation methodology incorporates a suite of metrics for various aspects that determine the final performance. This sequence of metrics will start with the Recognition Accuracy (RA) metric, evaluating the recognition accuracy of the problems in the visual prompts, i.e., if the LVLMs can recognize what are in the questions. Among those questions that an LVLM can accurately recognize the input, we then use the Instruction-following Effective Rate (ER) metric, calculating the average chances that an LVLM cannot yield an output compatible with the rule-based answer parser. Finally, the Aggregated Accuracy (AA) will calculate the weighted accuracy of LVLMs' answers correctness among those recognizable and parsable scenarios, aggregated with RA and ER. Figure 4 shows the sequence and relationship of these metrics.

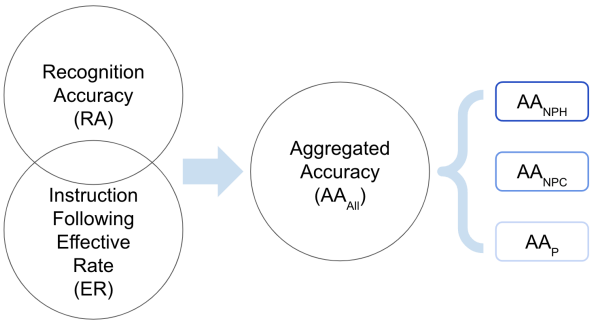

Figure 4: Metrics Sequence

Recognition Accuracy (RA) Recognition Accuracy evaluates the LVLMs' capability to accurately interpret the visual information presented in the prompts. This metric is crucial as it forms the foundation for any subsequent reasoning or decision-making processes the model undertakes. It is quantified by the ratio of prompts that are correctly recognized by the model to the total number of prompts presented, and is defined as:

$$RA = \frac{\sum_{i=1}^{N} C_i}{N}$$

where C_i is a binary indicator of correct recognition (1 if the ith prompt is correctly recognized, 0 otherwise), and N is the total number of prompts.

Instruction-following Effective Rate (ER) Instruction-following Effective Rate measures the average likelihood that an LVLM's response adhere to the expected output format, thus being compatible with a rule-based answer parser. This metric is crucial for gauging the models' reliability in producing solutions complying to standard parsing:

$$ER = \frac{\sum_{i=1}^{N} F_i}{N}$$

where F_i is parsing Instruction-following status for the ith problem, where $F_i = 1$ if the parsing task successes and $F_i = 0$ otherwise, and N is the total number of problems.

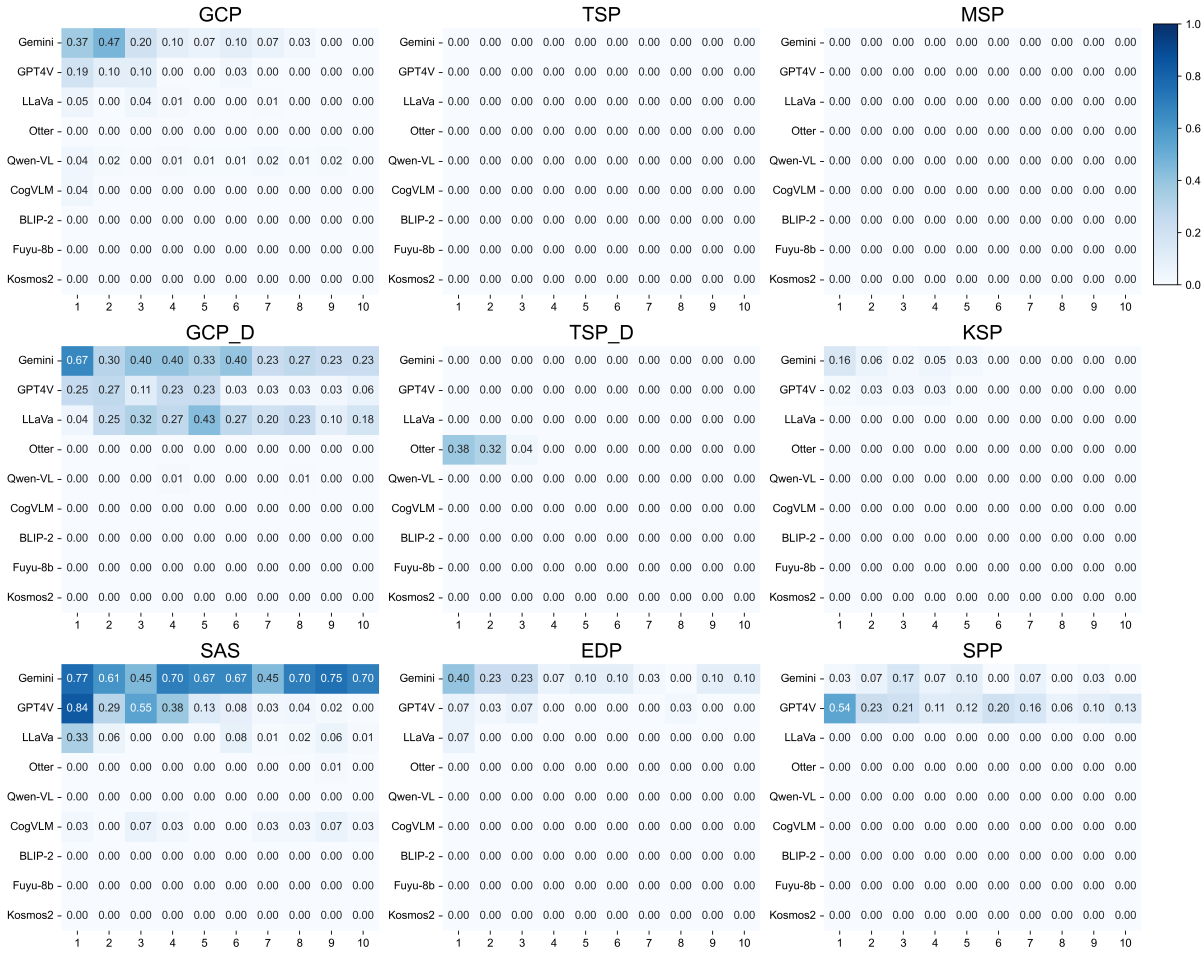

Figure 5: Reasoning abilities across models, complexity levels, and difficulty levels

Aggregated Accuracy (AA) Aggregated Accuracy takes into account the correct recognition of prompts and the successful parsing of responses to evaluate the correctness of LVLMs' answers. This metric is a weighted measure that integrates the Recognition Accuracy and the inverse of the Failure Rate:

$$AA = \frac{\sum_{i=1}^{N}(w_i \times A_i' \times RA_i \times (ER_i))}{\sum_{i=1}^{N} w_i}$$

where w_i represents the difficulty weight for the ith level, A_i' represents the accuracy in recognizable and parsable questions at i^{th} level, RA_i is the Recognition Accuracy for the ith level, and ER_i is the Instruction-following Effectiveness Rate for the ith level. **The Aggregated Accuracy metric will be used as the main evaluation metric for the overall performance and the complexity-specific submetrics**, demoted as AA or AA_{All}, AA_{NPH}, AA_{NPC}, and AA_P.

5 Results

5.1 Reasoning Abilities of LVLMs

Close and Open Source Models As Figure 8 indicates, the comparison between close source and open source LVLMs is quite stark, with close source models exhibiting superior performance in all tasks, irrespective of complexity class. Specifically, demonstrated in Figure 5, the Gemini-pro model consistently outperforms other models, including GPT-4V, across the benchmark. A detailed inspection of the heatmap reveals that Gemini-pro maintains a higher aggregated accuracy and a more effective Instruction-following rate than GPT-4V, which suggests that the proprietary nature in the Gemini-pro model may contribute to enhanced reasoning performance. This is especially evident in the NP-hard category, where Gemini-pro shows remarkable resilience in reasoning accuracy compared to GPT-4V, which exhibits a notable decline. We also provide the full results of aspect-based metrics for LVLMs in Figure 1.

Complexity Classes As Figure 5 indicates, the reasoning performance are inversely proportional to the complexity of the tasks. On simpler P problems, these models show commendable performance. However, as the complexity escalates to NP-complete and further to NP-hard problems, a clear and expected downtrend in reasoning performance is observed. This trend is consistent across all LVLMs, signifying a universal challenge in tackling higher-order complexity with current multimodal reasoning architectures. In addition to the overall trend, there are two notable findings. First, none of the nine models are capable of solving TSP and MSP. Second, the Otter model performs unexpectedly well on the TSP-D problem – all other models, including the close source ones are unable to solve any of the TSP-D problems.

Task Difficulties When focusing on individual reasoning tasks and considering models that can at least address the simplest questions, we notice a degradation in performance in correlation with increasing question difficulty. As Figure 5 shows, models like Gemini-pro display a high success rate on easier questions within tasks like GCP, but this success rate diminishes as the difficulty level increases. This pattern suggests that even the most capable LVLMs struggle to maintain their reasoning prowess as task difficulty intensifies.

5.2 Impact of Vision and Text Input

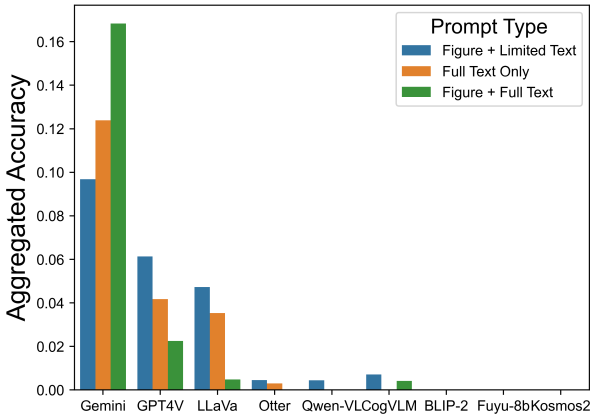

Figure 6: Reasoning abilities across prompt types

In evaluating the effect of different combinations of visual and textual inputs on the reasoning abilities of LVLMs, we observe varied responses across the models. Figure 6 indicates that a significant majority of the models, including both close source models like GPT-4V and open source models such as LLaVa, exhibit the highest levels of reasoning accuracy with the Default Setup, which pairs figures with limited instructional text (figure+limited_text). This suggests that their textual understanding is relatively weaker compared to their visual understanding, as the use of images to represent the question seems to be more effective than textual representation. Furthermore, when combining figure with full text, the models' performance does not improve and instead decreases, indicating that the addition of excessive textual information may hinder their reasoning abilities. Gemini-pro model stands out distinctly by demonstrating good reasoning performance with both the Text-only Setup (full_text_only) and the Vision-rich-text Setup (figure+full_text). It's also noteworthy that while the addition of full textual with image context to the visual prompts (Vision-rich-text Setup) improves Gemini-pro's performance, indicating that the integration of visual and textual information is indeed helpful for Gemini-pro.

These findings reveal that the integration of multimodal data processing in LVLMs does not necessarily lead to improved reasoning capabilities. The interplay between visual and textual inputs is not uniform across LVLMs and that certain models may be specially attuned to leverage text, visuals, or a combination of both to maximize their reasoning proficiency. It underscores the importance of prompt design in LVLM performance and calls for further investigation into how different models process multimodal information.

5.3 Comparison between LVLMs and LLMs on Reasoning

Figure 7 and Figure 8 provide a direct comparison between the current top-performing models in LVLMs and LLMs, including both closed-source and open-source models. In LLMs, we selected the models GPT-4-turbo, GPT-3.5-turbo, Claude-2, Claude-instant, PaLM-2, MPT-30b, Mistral-7b, Phi-1.5, Phi-2, Qwen-14b, and Yi-34b (Fan et al., 2023a). The evaluation is conducted on the exact same test dataset with the same questions, with the only difference being that the input prompt for LLMs are textual descriptions while the input prompt for LVLMs are the visual counterparts.

Figure 7. a presents the pure reasoning ability and Figure 7. b presents the instruction-following ability of various LVLMs, with the best performing ones being Gemini-pro and GPT-4V. Figure 8. a

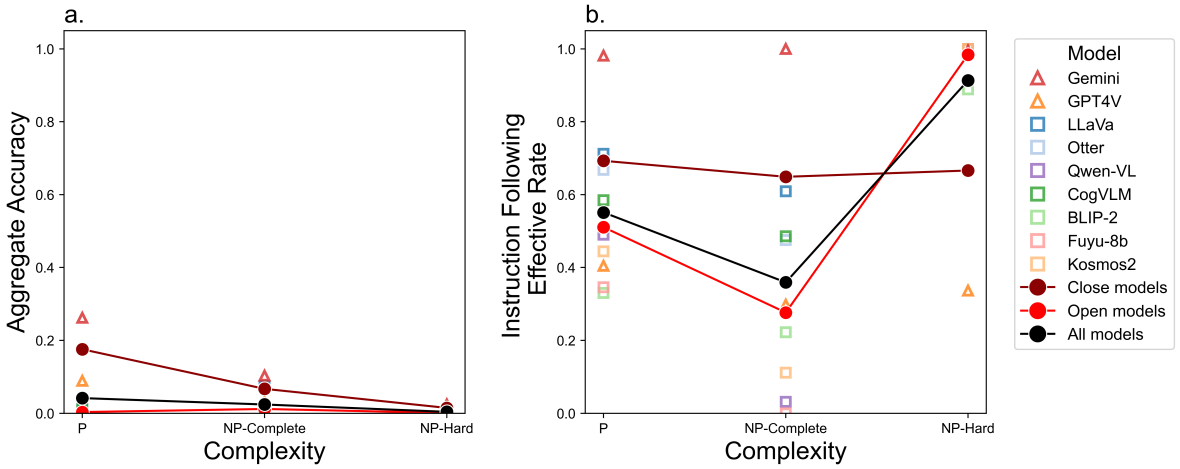

Figure 7: LVLM: a. Reasoning abilities performance excluding the effects of recognition and instruction following on figure+limited_text representation of questions. b. Instruction-following effective rate

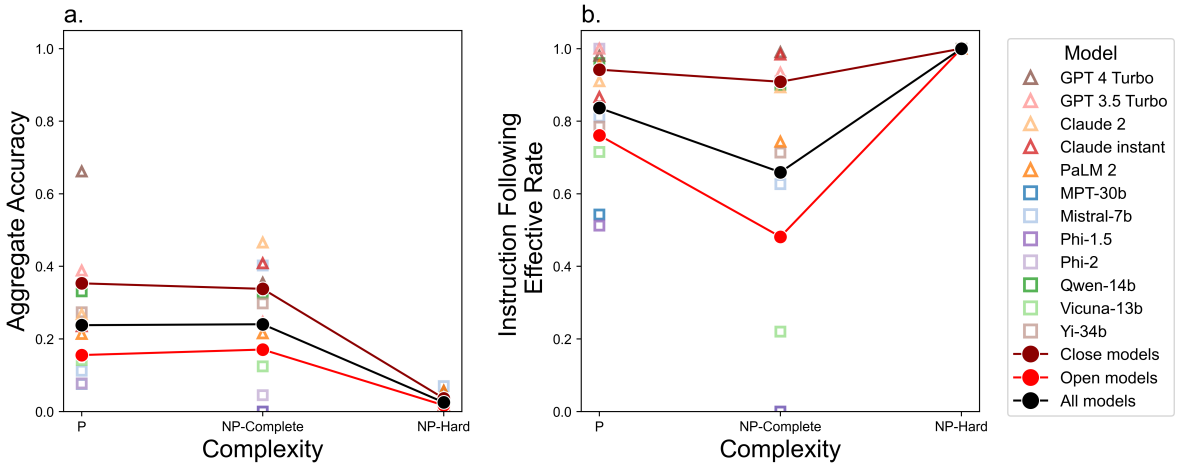

Figure 8: LLM: a. Reasoning abilities performance excluding the effect of instruction following on pure textual description of questions. b. Instruction-following effective rate

and b present those of LLMs with the best performing one being GPT-4-turbo. A cross comparison of these figures reveals that the reasoning ability of LVLMs generally lags behind that of LLMs. Specifically, the aggregated accuracy in P problems is approximately 0.4 in LLMs but only 0.2 in LVLMs. Furthermore, the decrease in aggregated accuracy is more pronounced in LVLMs compared to LLMs with the increase of computational complexity: While LLMs maintain a relatively consistent weighted accuracy in NP-complete and P problems, LVLMs' weighted accuracy decreases dramatically by half on average, from approximately 0.2 to 0.1. These findings highlight the need for further improvement in reasoning abilities of LVLMs.

6 Conclusion

In this paper, we propose the benchmark NPHard-Eval4V, a dynamic benchmark that scrutinizes the reasoning capabilities of LVLMs against the backdrop of computational complexity. The benchmark can be updated monthly to avoid hacking.

The experimental results show that the reasoning ability of LVLMs in general are still much weaker than that of LLMs. The experiments on the impact of vision and text input also suggest that the performance of LVLMs is contingent upon not only the complexity of the task but also the nature of input they are provided.

We have observed that while most models, including close source models such as GPT-4V and open source ones such as LLaVa, exhibit optimal performance with the Default Setup (figure with limited instructional text), it is the Gemini-pro model that stands out, showing a marked improvement in reasoning ability when provided with text-only and vision-rich-text (figure with full textual descriptions) prompts.

Limitations

Inherent Bias in Model Comparisons The benchmarking process inherently favors models like Gemini-pro over others due to the varied architectural strengths, which may not be fully captured in a uniform evaluation setup. Despite our attempts to level the playing field, certain models are naturally more adept at specific types of reasoning tasks due to their difference in training data, which could skew the overall comparative analysis. For example, Gemini's superior performance in text-only setups may reflect an underlying bias in the benchmark towards linguistic reasoning over multimodal reasoning.

Limited Scope of Reasoning Tasks Our benchmark only scratches the surface of the vast landscape of reasoning. The tasks chosen, while diverse, cannot possibly encompass the full spectrum of reasoning abilities that LVLMs might possess. There is a risk that the benchmark may overrepresent certain reasoning styles while underrepresenting others, leading to an incomplete picture of models' true reasoning capacities.

Prompt Dependence The study highlights the variability in model performance based on prompt type, suggesting a significant dependence on how questions are framed. This raises concerns about the models' robustness and their ability to generalize across different input conditions. It is unclear whether the observed performance reflects true reasoning ability or an artifact of the models' sensitivity to specific prompt structures.

Multimodal Integration Challenges The weaker performance of LVLMs in comparison to LLMs suggests that the integration of multimodal data may not always be beneficial for reasoning tasks. This observation implies limitations in how current models process and integrate multimodal information, potentially indicating a misalignment between model architecture and task requirements.

Ethics

Ethical Considerations In this work, we introduce the **NPHardEval4V** benchmark, designed to rigorously evaluate the reasoning capabilities of Large Vision-Language Models (LVLMs) through algorithmic tasks ranging in complexity from polynomial time to NP-hard problems. While the benchmark aims to enhance our understanding of LVLMs, we acknowledge potential ethical concerns and have taken steps to address them.

Fair and Unbiased Evaluation We have ensured that the NPHardEval4V benchmark provides an impartial and equitable evaluation of LVLMs, regardless of their development source or specific architecture. Our careful selection of tasks, spanning various complexity classes, is designed to provide a broad assessment of reasoning abilities without bias toward any particular model type or development approach. We remain committed to objective evaluation practices and transparency in scoring methods to avoid any unfair advantages or biases.

Data Privacy and Security Although NPHardEval4V does not involve human subjects or personal data, we prioritize security and privacy in all aspects of the benchmark. The dataset used consists solely of algorithmic problems, eliminating concerns around personally identifiable information (PII). Additionally, the dynamic update mechanism regularly refreshes the dataset, retaining only the latest three months of data to mitigate any potential risks of model exploitation or gaming the benchmark system.

Transparency and Reproducibility To promote reproducibility and transparency, the benchmark's design, dataset, and code are well-documented and made available for the research community. By providing detailed implementation guidelines, we ensure that our methods are replicable, fostering further development and collaborative research. The dynamic update mechanism and all associated benchmarks are accessible via an open, anonymous repository, allowing for continued advancements in evaluating LVLM reasoning abilities.

Responsible Development and Use of AI Our research is rooted in the principle of advancing AI technology responsibly. The NPHardEval4V benchmark is designed not only to push the frontiers of LVLM capabilities but also to guide the development of AI systems that are beneficial and aligned with ethical standards. We aim to foster models that are capable of advanced reasoning, with potential applications that contribute positively to society while mitigating risks of misuse.

Prevention of Overfitting A critical issue in AI benchmarking is the risk of models being tailored too specifically to a static dataset, leading to inflated and inaccurate evaluations of their true abil-

ities. To counter this, the dynamic update mechanism ensures that the benchmark continuously evolves, preventing overfitting and promoting genuine advancements in LVLM performance. This approach helps maintain the integrity of model evaluation and encourages authentic improvements over superficial gains.

Accessibility and Inclusivity We are committed to making the NPHardEval4V benchmark widely accessible to researchers and developers across the AI community. Clear documentation and usage guidelines have been provided to ensure that users from diverse backgrounds and experience levels can effectively engage with the benchmark. This inclusivity promotes a rich, collaborative environment for AI research, encouraging contributions and insights from a wide array of perspectives.

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.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. 2023. Introducing our multimodal models.

Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. 2023. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. *arXiv preprint arXiv:2311.02782*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*.

Shan Chen, Mingye Gao, Kuleen Sasse, Thomas Hartvigsen, Brian Anthony, Lizhou Fan, Hugo Aerts, Jack Gallifant, and Danielle Bitterman. 2024. Wait, but tylenol is acetaminophen... investigating and improving language models' ability to resist requests for misinformation.

Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. 2023. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. *arXiv preprint arXiv:2311.00571*.

Lizhou Fan, Wenyue Hua, Lingyao Li, Haoyang Ling, and Yongfeng Zhang. 2023a. Nphardeval: Dynamic benchmark on reasoning ability of large language models via complexity classes. *arXiv preprint arXiv:2312.14890*.

Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. 2023b. A bibliometric review of large language models research from 2017 to 2023. *arXiv preprint arXiv:2304.02020*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2023a. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Chaoyou Fu, Renrui Zhang, Haojia Lin, Zihan Wang, Timin Gao, Yongdong Luo, Yubo Huang, Zhengye Zhang, Longtian Qiu, Gaoxiang Ye, et al. 2023b. A challenger to gpt-4v? early explorations of gemini in visual expertise. *arXiv preprint arXiv:2312.12436*.

Ran Gong, Qiuyuan Huang, Xiaojian Ma, Hoi Vo, Zane Durante, Yusuke Noda, Zilong Zheng, Song-Chun Zhu, Demetri Terzopoulos, Li Fei-Fei, et al. 2023. Mindagent: Emergent gaming interaction. *arXiv preprint arXiv:2309.09971*.

Google. 2023. Meet the first version of gemini— our most capable ai model.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Jiaxian Guo, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Boyang Li, Dacheng Tao, and Steven Hoi. 2023. From images to textual prompts: Zero-shot visual question answering with frozen large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10867–10877.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. 2023. Grounding language models to images for multimodal inputs and outputs.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

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

Fuxiao Liu, Tianrui Guan, Zongxia Li, Lichang Chen, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2023a. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multi-modality models. *arXiv preprint arXiv:2310.14566*.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023b. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Yuliang Liu, Zhang Li, Hongliang Li, Wenwen Yu, Mingxin Huang, Dezhi Peng, Mingyu Liu, Mingrui Chen, Chunyuan Li, Lianwen Jin, et al. 2023c. On the hidden mystery of ocr in large multimodal models. *arXiv preprint arXiv:2305.07895*.

Zhaoyang Liu, Zeqiang Lai, Zhangwei Gao, Erfei Cui, Zhiheng Li, Xizhou Zhu, Lewei Lu, Qifeng Chen, Yu Qiao, Jifeng Dai, et al. 2023d. Controllm: Augment language models with tools by searching on graphs. *arXiv preprint arXiv:2310.17796*.

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. 2023. Gpt-4v(ision) system card.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. 2023. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. 2023. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *arXiv preprint arXiv:2306.09265*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*.

Fei Yu, Hongbo Zhang, and Benyou Wang. 2023. Nature language reasoning, a survey. *arXiv preprint arXiv:2303.14725*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. 2024. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36.

A Benchmark Details: Models and Tasks

Table 1 shows the nince VLMs we assessed in the NPHardEval4V benchmark.

Model	n Parameters (B)	Access
GPT-4V (OpenAI, 2023)	Unknown	Close source
Gemini-pro 1.0 Pro (Google, 2023)	Unknown	Close source
CogVLM (Wang et al., 2023)	17	Open source
LLaVA-1.5-13B (Liu et al., 2024)	13	Open source
BLIP-2 FLAN-T5-XXL (Li et al., 2023c)	11	Open source
Fuyu-8B (Bavishi et al., 2023)	8	Open source
Otter (Li et al., 2023a)	8	Open source
Qwen-VL-7B (Bai et al., 2023)	7	Open source
Kosmos2 (Peng et al., 2023)	1.5	Open source

Table 1: LVLMs metadata

Table 2 shows the detailed task (question) choices. The questions are the same set as the NPHardEval benchmark (Fan et al., 2023a).

B Recognition Accuracy

Recognition is an important preprocess that disentangles input processing's impact on reasoning ability of LVLMs. In general, close source models outperforms open source counterparts in recognition. It is also important to notice irrelevance between complexity of tasks and recognition accuracy.

Several close source models, however, can obtain similar or even better recognition ability than the open source ones. For instance, CogVLM constantly reaches full recognition rate (1.0) on SAS and KSP problems. LLaVa, Otter, and QWen-VL also have different extents of outstanding recognition abilities on different tasks. In addition to further showing LVLMs' variation in tasks regarding recognition ability, this result also reveals the individual strengths of different LVLMs, possibly due to different training data and techniques.

C Prompt Examples

C.1 SPP Prompt Example for the Default Setup

Below is an example for an SPP problem:

```
The Shortest Path Problem (SPP) involves finding the shortest path between two nodes in a weighted graph.
You need to find the shortest path between node 0 and node 3 in a graph. The graph's edges and their weights are given.
Please provide the shortest path from 0 to 3 and its total distance. Offer a concise step-by-step explanation of your reasoning process. Aim for brevity and clarity in your response.
Your output should be enclosed within <root></root> tags. Include your reasoning in <reasoning></reasoning> tags and the final path and total distance in <final_answer></final_answer> tags, like <final_answer>{'Path': 'START->...->END', 'TotalDistance': 'INT_TOTAL_DISTANCE'}</final_answer>.
```

C.2 SPP Prompt Example for the Text-only Setup

In the SPP example corresponding to the example in Appendix C.1, the pure textual prompt is shown as below:

```
The Shortest Path Problem (SPP) involves finding the shortest path between two nodes in a weighted graph.
You need to find the shortest path between node 0 and node 3 in a graph. The graph's edges and their weights are given.
Please provide the shortest path from 0 to 3 and its total distance. Offer a concise step-by-step explanation of your reasoning process. Aim for brevity and clarity in your response.
Your output should be enclosed within <root></root> tags. Include your reasoning in <reasoning></reasoning> tags and the final path and total distance in <final_answer></final_answer> tags, like <final_answer>{'Path': 'START->...->END', 'TotalDistance': 'INT_TOTAL_DISTANCE'}</final_answer>.

In this graph:
The distance between node 0 and 1 is 1,
the distance between node 1 and 3 is 9,
the distance between node 0 and 3 is 5,
the distance between node 0 and 2 is 2,
the distance between node 2 and 3 is 4.
```

Complexity Class	Task
NP Hard (Most Complex)	Graph Coloring Problem Optimization Version (GCP) Traveling Salesman Problem Optimization Version (TSP) Meeting Scheduling Problem (MSP)
NP Complete	Knapsack Problem (KSP) Traveling Salesman Problem Decision Version (TSP-D) Graph Coloring Problem Decision Version (GCP-D)
P (Least Complex)	Shortest Path Problem (SPP) Edit Distance Problem (EDP) Sorted Array Search (SAS)

Table 2: Complexity classes and tasks

D Research Outlook

Longitudinal Learning Studies Investigating models' learning curves over extended periods could offer valuable insights into their potential for growth and adaptation. Long-term studies that track how models evolve with additional training data or through incremental learning can provide a deeper understanding of the underlying mechanisms that drive progress in LVLMs and identify the key factors that contribute to long-term success.

Expanding Reasoning Taxonomies Research should continue to expand the taxonomy of reasoning tasks within benchmarks to ensure a comprehensive evaluation of LVLMs. This includes not only increasing the variety of tasks but also ensuring that they are representative of the diverse ways in which reasoning is applied in real-world contexts. Developing benchmarks that can simulate more complex, real-life reasoning scenarios will be crucial for advancing the field.

Harmonizing Model Evolution with Benchmarks Balancing the dynamic nature of benchmarks with the pace of model development is essential. Researchers could explore introducing phased or tiered updates to benchmarks that allow for more strategic model improvements, rather than continuous, potentially disjointed, monthly updates. This approach could better align the benchmark evolution with the natural R&D cycles in AI development.

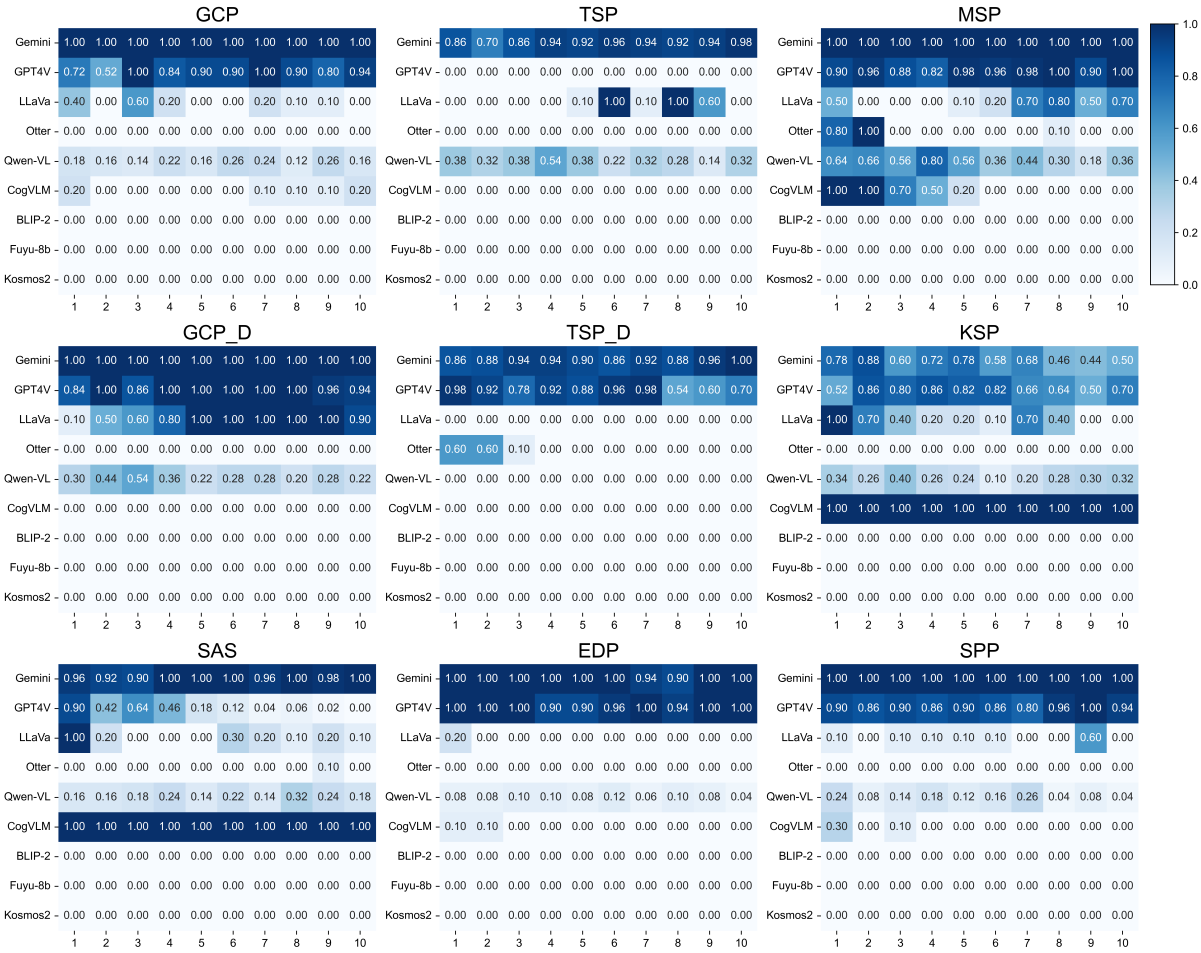

Figure 9: Prompt recognition across models, complexity levels, and difficulty levels

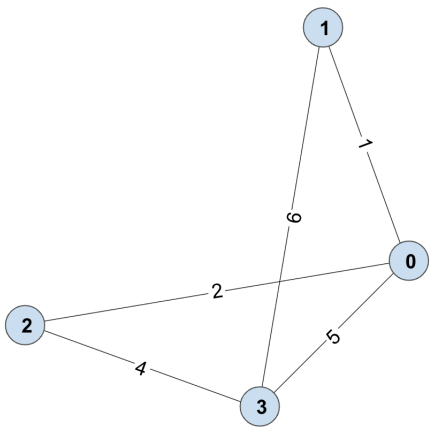

Figure 10: Image of an Short Path Problem example