Mixed Signals: Decoding VLMs' Reasoning and Underlying Bias in Vision-Language Conflict

Anonymous ACL submission

Abstract

Vision-language models (VLMs) have demonstrated impressive performance by effectively integrating visual and textual information to solve complex tasks. However, it is not clear how these models reason over the visual and textual data together, nor how the flow of information between modalities is structured. In this paper, we examine how VLMs reason by analyzing their biases when confronted with scenarios that present conflicting image and text cues—a common occurrence in real-world applications. To uncover the extent and nature of these biases, we build upon existing benchmarks to create five datasets containing mismatched image-text pairs, covering topics in mathematics, science, and visual descriptions. Our analysis shows that VLMs favor text in simpler queries but shift toward images as query complexity increases. This bias correlates with model scale, with the difference between the percentage of image- and text-preferred responses ranging from +56.8% (image favored) to -74.4% (text favored), depending on the task and model. In addition, we explore three mitigation strategies: simple prompt modifications, modifications that explicitly instruct models on how to handle conflicting information (akin to chain-of-thought prompting), and a task decomposition strategy that analyzes each modality separately before combining their results. Our findings indicate that the effectiveness of these strategies in identifying and mitigating bias varies significantly and is closely linked to the model's overall performance on the task and the specific modality in question. We will release our dataset and code.

1 Introduction

Vision-language models (VLMs) have rapidly advanced the field of artificial intelligence by effectively combining visual and textual data to achieve state-of-the-art performance on a variety of tasks

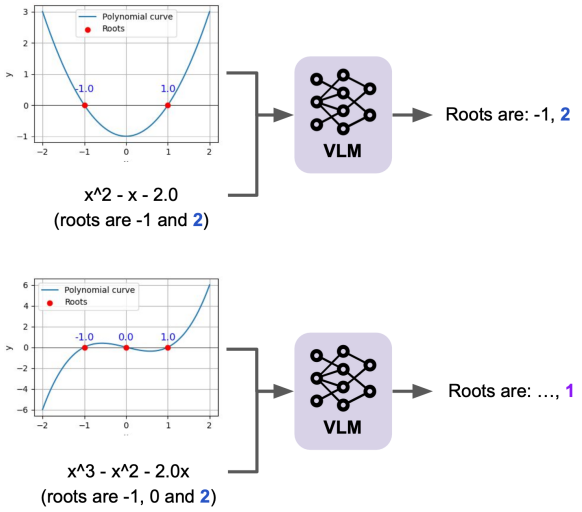


Figure 1: We investigate VLMs' bias toward text versus image inputs when mismatches occur between the modalities. Our observations reveal that this bias heavily depends on the task's and sample's difficulty. For example, while the model relies on textual representations to compute the roots of a degree-2 polynomial, increasing the degree to 3 shifts the reliance more toward the visual representation of the function.

(Zhou et al., 2022; Liu et al., 2023b; Zhang et al., 2024; Li et al., 2025). However, it remains unclear how these models integrate and reason over information from multiple modalities—a capability that becomes increasingly important in applications such as retrieval-augmented generation (RAG) (Yu et al., 2024; Yuan et al., 2024) and multi-agent systems (Ghafarollahi and Buehler, 2024; Jiang et al., 2024), where data can be sourced from diverse modalities.

To understand how VLMs reason over multiple modalities, one approach is to investigate their behavior when confronted with conflicting information. Moreover, it is important to determine whether these models exhibit a bias toward one modality over the other, when exposed to deliberately mismatched image-text pairs (see Figure 1). For example, consider a scenario in healthcare

where a chest X-ray image shows no signs of pneumonia, yet the accompanying text erroneously describes clear evidence of the disease. If the model disproportionately relies on the text, it may incorrectly diagnose pneumonia, potentially leading to inappropriate treatment recommendations and severe patient harm.

In this paper, we investigate VLMs' biases in the presence of conflicting multimodal cues by building upon existing datasets to create five novel benchmarks (Liu et al., 2023a; Fu et al., 2024). Each benchmark consists of mismatched textual and image cues, with every sample assigned a complexity score to study the impact of sample difficulty on the models bias. These benchmarks span various topics, including mathematical reasoning, science, and visual descriptions. To generate the mismatched pairs, we begin with aligned image-text pairs from the original datasets and then employ a combination of rule-based methods and manual modifications to alter the textual content, thereby introducing conflicting information.

We adopt five state-of-the-art VLMs spanning a diverse range of scales and conduct experiments on our constructed benchmarks. Our analysis reveals that when presented with conflicting information, VLMs tend to favor textual data over images in simpler scenarios. However, as query complexity increases, the models shift their bias toward the modality they perceive to be simpler—in many cases, this is the image. Furthermore, we observe that the extent of this bias correlates with the model's scale and strength. Finally, our findings indicate that this modality preference arises partly from the models' perception of task difficulty and partly from other internal biases, and is strongly linked to LLMs performance with different modalities as the input on a given task.

Observing the significant degree of modality bias in our experiments, we set out to explore several mitigation strategies aimed at reducing this issue. We consider three approaches to address modality bias: simple prompt modifications, introducing explicit instructions similar to chain-of-thought prompting, and a decomposition approach that analyzes each modality separately before combining outputs. Our experiments show that each approach exhibits a diverse impact based on the task, the model, and—most importantly—the model's performance when given different modalities as input. In scenarios where the model shows strong performance with both modalities (or even one modality) as the input, at least one of explored mitigation strategies demonstrate reasonable performance.

2 Conflicting Modalities

In this work, our goal is to understand how VLMs reason over multiple modalities when they are presented with conflicting information. We simulate realistic scenarios in which the visual and textual inputs contradict one another, forcing the models to weigh and integrate disparate cues. In this section, we outline the problem setting and describe the methodology used to construct our benchmarks. We provide the prompts used in creating the benchmark and the data statistics of created benchmarks in the Appendix.

2.1 Problem Statement

To evaluate how vision-language models (VLMs) handle conflicting multimodal cues, we formulate the following experimental setting. Assume we have a query Q, an image I, and a corresponding textual description T each yield a consistent answer A. Each query is associated with a complexity score c that reflects the difficulty of the reasoning task. We then create a modified textual description T' that intentionally provides an answer A' different from A.

In an ideal scenario, a VLM would detect the discrepancy between I and T' and indicate a conflict rather than committing to either A or A'. However, if the model outputs an answer that aligns with one of the modalities (A or A'), this behavior is considered a bias toward that modality. We can quantify this bias by defining the bias metric B as:

$$B = f(|A|, |A'|),$$

where $|\cdot|$ represents the number of times the model adheres to the response from a specific modality, and f is a function measuring the difference between the inputs, which can be as simple as computing their ratio. This metric captures the degree to which the model's output favors certain modality. In our experiments, we define B as the difference between the percentage of image-favored responses and the percentage of text-favored responses.

2.2 Benchmarking

To systematically study VLM's biases, we build upon data from two existing datasets—VSR (Liu et al., 2023a) and Isobench (Fu et al., 2024)—and create five distinct benchmarks featuring mismatched image-text pairs.

Graph Connectivity: For this benchmark, we start with graph samples sourced from Isobench. We manually modify the adjacency matrices in a minimal manner to alter the connectivity between target nodes. When the original graph is unconnected, we ensure that our modifications do not simply connect the two target nodes directly, preserving the underlying structure while ensuring certain level of difficulty. We approximate the complexity of each sample in this task by the number of edges in its corresponding graph.

Function Convexity: Using samples from Isobench, we generate conflicting pairs by altering the functions expressions. Specifically, we multiply each coefficient of the function by minus one, creating a scenario where the textual description of the function's convexity contradicts the visual representation. We approximate the complexity of each sample in this task based on the number of characters in the textual representation of the mathematical expression of the function.

Polynomial Roots Calculation: In this benchmark, we generate polynomials of degrees 1 through 4 (since they have closed form solution) by randomly selecting the roots within the range of –10 to 10 and use them to construct both the polynomial expressions and their corresponding visual representations. To create a conflicting pair, we randomly alter one of the roots by replacing it with a different value from the same range, resulting in a discrepancy between the textual and visual depictions of the polynomial. The degree of each polynomial serves as a proxy for task complexity.

Physics and Chemistry Questions: For this task, we leverage samples from Isobench and manually alter the textual descriptions of physics and chemistry problems. The modifications are designed so that the answer derived from the text would differ from that suggested by the visual cues, pointing to a choice other that the initial answer from the provided multiple-choices. Additionally, we filter out any questions that could be correctly answered by relying solely on the question, ensuring that the conflict between image and text is both meaningful and challenging. We manually assign an "easy" or "hard" label to each sample to reflect its complexity.

Visual Description: For this benchmark, we first select samples from the VSR dataset in which the original statement accurately reflects the content of the image. We then use GPT-4o mini to generate an extended description of the image centered around the given statement. By manually identifying the opposite of spatial relationships studied in the VSR dataset, we replace the original relation with its opposite (in the extended description), thereby creating a mismatched pair. We analyze the per-relation breakdown (based on the original input relation) of VLMs performance as a proxy for task complexity.

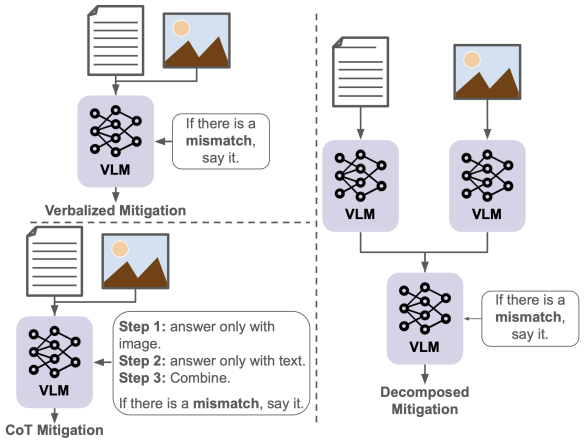


Figure 2: We investigate the impact of three mitigation strategies—Verbalized, CoT, and Decomposed—on identifying mismatches in the input modalities.

3 Mitigation Strategies

Our experiments indicate that VLMs exhibit a notable bias when handling conflicting multimodal information (see section 4.2). To address this challenge, we propose three distinct mitigation strategies, as summarized in Figure 2 with the adopted prompts provided in the Appendix. These strategies aim to detect and mitigate the bias by altering the models' processing of visual and textual inputs, thereby encouraging a more balanced integration of information. The strategies are as follows:

Verbalized Mitigation: Directly prompting the model to identify and report any mismatches or contradictions between the modalities. Instead of asking the models to provide only the answer, we additionally instruct them to indicate a mismatch if a discrepancy is detected. This explicit acknowledgment of conflict helps prevent the model from defaulting to one modality.

CoT Mitigation: Taking inspiration from chain-of-thought prompting, in this method, the model is guided through a three-step process. First, we instruct the model to process the image input alone;

second, evaluate the textual description independently; and third, combine the outputs from both modalities. We further instruct the model to extract all relevant information from each modality, particularly when a single modality does not suffice to solve the task. Finally, the model is instructed to compare the two results and either provide an answer or indicate if a mismatch is detected.

Decomposed Mitigation: In this method, we employ a multi-stage approach wherein the VLM is run three separate times. The first run solve the task using only the image input, the second only use the text input, and the third combines the outputs from the previous two runs. In the final stage, the model is specifically instructed to highlight any inconsistencies between the outputs, and if a discrepancy is detected, to report a mismatch.

4 Experiments

In here, we first examine the extent of bias in VLMs using our created benchmarks. Next, we examine how bias correlates with the models' perceived problem difficulty and conduct an error analysis of their performance. Finally, we explore the impact of various mitigation strategies in reducing the bias.

4.1 Experimental Details

For our experiments, we evaluate five state-of-the-art vision-language models: Qwen2-VL-7B-Instruct and Qwen2-VL-72B-Instruct (Wang et al., 2024b), Llam-3.2-90B-vision-instruct (Grattafiori et al., 2024), GPT-4o mini, and GPT-4o (Hurst et al., 2024). These models have been selected for their diverse architectures and multimodal processing capabilities, enabling us to investigate how factors such as model scale, training methodology, and fusion strategies affect the integration of conflicting information. Moreover, we rely on accuracy and F1 scores as evaluation metrics. To determine the models' preferred modality for each sample, we identify the modality-specific label with which the output aligns. All prompts used with these models are provided in the Appendix.

4.2 Bias in VLMs

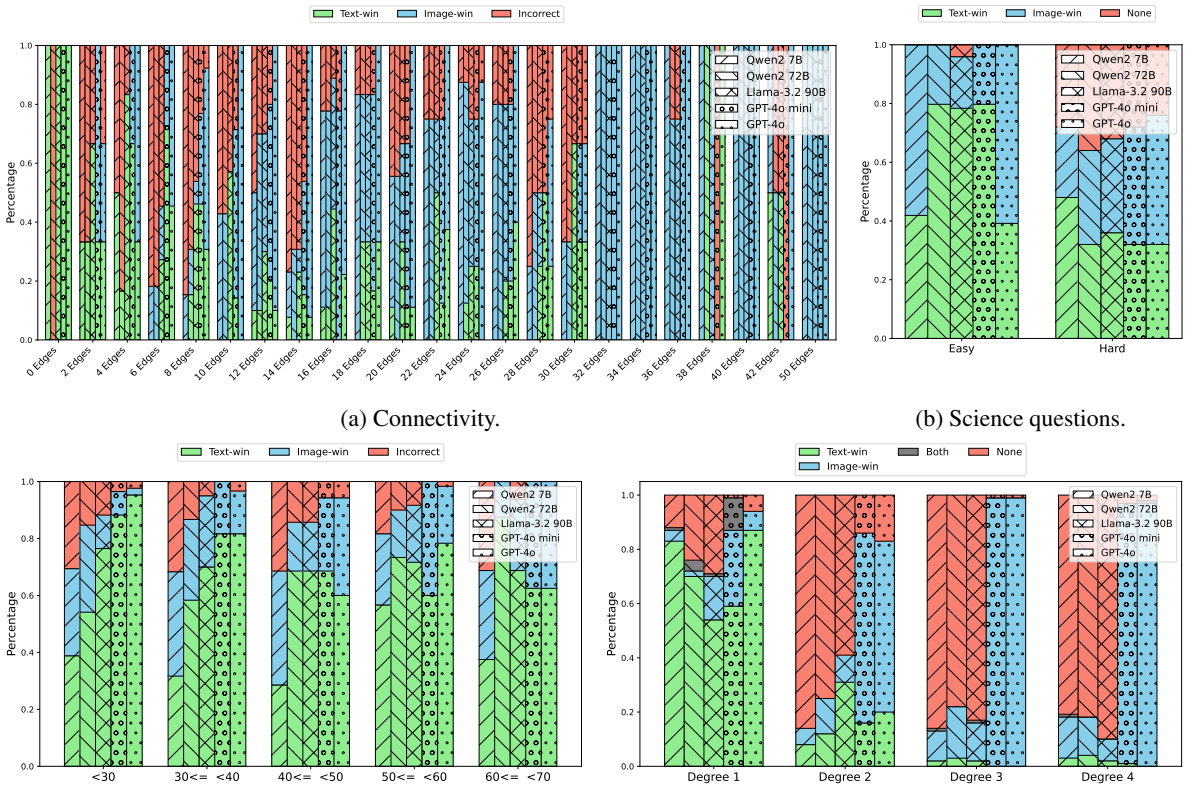
We measure the bias of VLMs toward text or image modalities using our created datasets. To isolate the impact of bias from the models mispredictions, we calculate the percentage of text versus image bias in mismatched inputs only for those samples where models produced correct predictions when both the image and text were aligned (i.e., the original inputs). Additionally, we report the percentage of incorrect predictions on the original, aligned inputs. We also present the accuracy of models' performance for solving each task when using only one modality or both, along with the corresponding B value for each task and model in the Appendix.

4.2.1 Bias in Mathematical Reasoning

Graph Connectivity: We present the distribution of modality preferences across graphs in Figure 3a, where the task is to determine whether two target nodes are connected. We categorize model performance by the number of edges in the graphs—a proxy for task difficulty. For simpler graphs with fewer edges, models tend to favor answers supported by textual input. However, as graph difficulty increases, there is a clear shift toward reliance on visual (image-based) information. Notably, for highly complex graphs (those with 32 edges or more), most models rely exclusively on image inputs to generate their responses. Moreover, while all models exhibit similar bias trends, Qwen models show a higher rate of incorrect predictions. Finally, based on the overall bias values reported in the Appendix (Table 4), we observe that, with the exception of Llama-3.2 90B, all other models generally favor images over text in this task.

Function Convexity: The bias distribution of VLMs in identifying the convexity of mathematical functions is depicted in Figure 3c. We approximate task difficulty based on the number of characters in each function's mathematical expression and divide the samples into five categories. The results indicate that for stronger models, as the difficulty increases, the textual bias decreases while the reliance on image cues increases. In contrast, for Qwen models, we observe a mixed impact of difficulty on modality bias. We suspect this inconsistency may be partly due to noise in our difficulty approximation method, as reflected by the higher number of incorrect predictions in some of the easier categories (those with fewer characters). Finally, based on the overall bias values reported in the Appendix (Table 4), we find that all models generally favor text over image in this task.

Polynomial Roots Calculation: The results of VLM bias in calculating the roots of polynomials are presented in Figure 3d. As shown, there is a clear shift in modality preference as the degree of the polynomial increases. For polynomials of



(a) Connectivity.

(b) Science questions.

(c) Convexity.

(d) Roots calculation.

Figure 3: The distribution of VLMs biases toward text versus image inputs.

degree 1, the models exhibit a strong bias toward textual input, indicating that simpler algebraic reasoning is more effectively grounded in the accompanying text. In some instances, a few models even correctly identify the mismatch by reporting both the image- and text-derived roots. However, as we move to polynomials of degree 2 and higher, this reliance on textual information diminishes substantially, with models increasingly favoring visual (image-based) inputs. This transition likely reflects the growing complexity of the problem space, where interpreting graphical representations becomes more advantageous—or even necessary—for accurate problem solving. The models' adaptive use of modalities suggests an emergent reasoning behavior, wherein they selectively leverage the most informative input source based on the task's complexity. Drawing on the overall bias values reported in the Appendix, we observe that OpenAI models tend to favor images, while other models generally favor text in this task.

4.2.2 Bias in Science Questions

In answering scientific questions, we observe a notable reliance on textual information for easier questions, suggesting that the model is more confident in leveraging text when the reasoning demands are relatively low (see Figure 3b). However, this reliance diminishes markedly for more challenging questions. Although support from image inputs appears to increase in these more challenging cases, a significant proportion of responses fall into the incorrect predictions. This trend suggests that, within the domain of science questions, the model struggles to produce a conclusive answer when faced with complex reasoning tasks that require effective integration of both text and image modalities. These findings potentially point to a limitation in the model's ability to reconcile multimodal information under higher cognitive load. The overall trend in the bias values indicates that, except for Qwen2 7B and GPT-4o—which favor images—all other models favor text in this task (see Table 4).

4.2.3 Bias in Visual Description

Investigating VLMs' biases toward text or image inputs using VSR samples, we observe distinct behaviors depending on the model and the specific spatial relation queried. For instance, GPT-4o exhibits a stronger bias toward textual data for spatial relations such as "perpendicular to", "facing away from","in the middle of", "parallel to", "on",

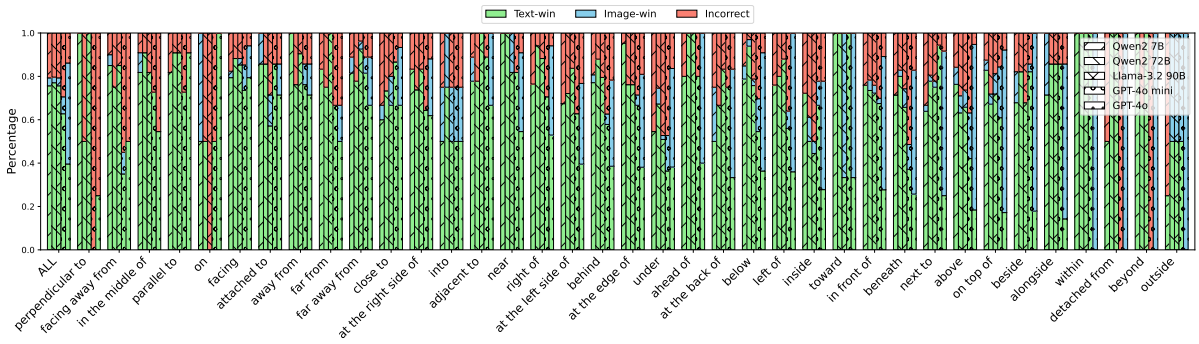


Figure 4: The distribution of VLM biases toward text versus image in the VSR-based dataset. We report the per-spatial relationship (base on the original input's relation) break down of performance.

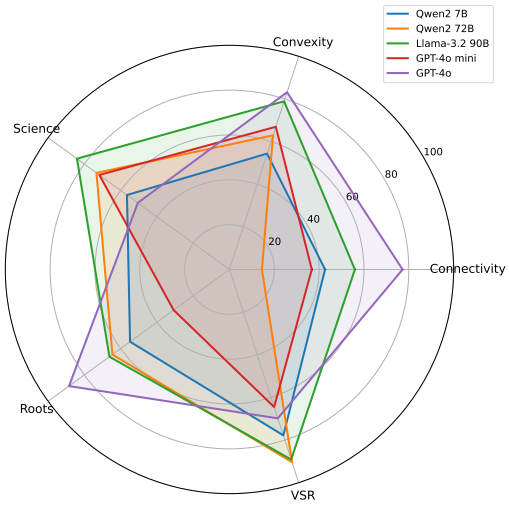


Figure 5: The accuracy of VLMs' internal perception of the simpler modality for solving the task is evaluated by comparing it to the actual modality each model relies on during problem solving.

"facing", and "attached to". In contrast, for relations like "outside", "beyond", "detached from", "within", "alongside", "beside", "on top of", and "above", the model shows a much greater bias toward image data, highlighting the strong impact of specific spatial relations on the VLMs modality preference. This variation is less pronounced in other models, although shifts in modality bias between text and image inputs are still evident across different spatial relations. We suspect that these differences largely reflect the training data, where the distribution of spatial relations varies across modalities. Based on the overall bias values reported in the Appendix (Table 4), we observe that, except for GPT-4o, all other models generally favor text over image in this task.

4.3 VLMs Bias and Their Perceived Sample Difficulty

In this section, we investigate the relationship between the models' inherent modality bias and their internal estimation of problem difficulty. Specifically, for each sample where the model provided a correct answer for initial matching inputs, we consider the modality it relied on (with mismatching inputs)—either text or image—as the gold label, representing the modality it implicitly deemed easier to use. Samples with incorrect initial answers are excluded to ensure that the gold label reflects a reliable internal assessment. Subsequently, for each sample from our datasets, given the mismatched pair, we asked the models to explicitly predict which modality (text or image) they believed would be easier to use in answering the query. The accuracy of these predictions, which is summarized in Figure 5, reveals trends similar to those observed in our earlier analysis. In particular, stronger models show a clear alignment between their perceived ease of using a given modality and the bias they exhibit when resolving a task with conflicting visual and textual information. This correlation underscores the role of internal difficulty estimation in driving modality bias, suggesting that models tend to favor the modality they internally assess as less challenging for solving the task.

In cases where we observe low accuracy—highly correlated with the models' overall task performance (see Table 3 in the Appendix)—we suspect that additional internal biases influence the models' modality preferences. Analysis of the explanations provided by the models reveals that, in most of those instances, they do not rely on the specific details of the sample but rather use the general task description to determine which modality is easier to process.

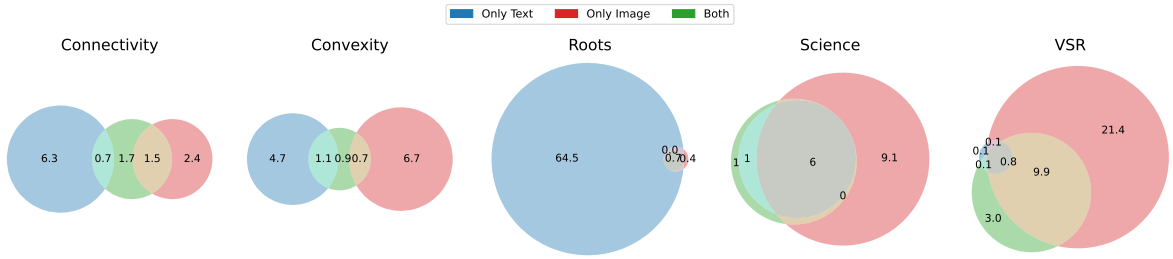


Figure 6: **Venn diagram of errors** displaying the percentage of mispredicted samples when GPT-4o is provided with only text, only image, and both text and image inputs.

4.4 VLMs Bias and The Models Performance

One possible approach to understanding the roots of modality bias in VLMs is through error analysis. By examining the correlation between model failures across different input modalities, we can gain deeper insights into the underlying mechanisms driving its decision-making. In Figure 6, we present a Venn diagram illustrating GPT-4o's error percentages under three different input settings: text-only, image-only, and combined matching initial text-image inputs (Venn diagrams of other models are provided in the Appendix). Analyzing the percentage of unique and overlapping errors (error independence) enables us to connect these failure modes to the biases observed earlier. For instance, if GPT-4o's errors unique to text-only and image-only inputs are resolved when both modalities are provided, we can conclude that the model effectively integrates the two modalities and reason over them. On the other hand, errors unique to a single modality that persist even when both text and image inputs are provided reveal a blind bias toward that modality, further underscoring the model's inability to effectively reason across both modalities. Consequently, a higher proportion of such cases signals a stronger bias in favor of that modality.

As shown, the number of unique-modality errors that persist after combining both text and image inputs is higher for images in the connectivity and VSR tasks, while it is higher for text in the convexity task, highlighting greater model bias toward those modalities (this aligns with the overall bias values reported in Table 4 in the Appendix). Moreover, in the root calculation and science tasks, the number of persistent errors appears similar for both text-only and image-only inputs. Beyond highlighting these modality-specific failure patterns, diagrams also underscore how task complexity influences model performance. Specifically, GPT-4o's errors tend to cluster in a single modality for the root calculation, science, and VSR, whereas for connectivity and convexity tasks, errors are more evenly distributed between modalities. This suggests that the complexity of solving these tasks is more balanced between the modalities.

4.5 Mitigating Bias in VLMs

To investigate the impact of our mitigation strategies, we calculate their accuracy in identifying mismatches over combination of initially matched and modified mismatched text-image pairs for each dataset (resulting in balanced evaluation sets). The results of bias mitigation strategies are summarized in Table 1 (the Qwen2 7B model was incapable of solving the task when the image was missing). Our findings demonstrate that if models achieve high accuracy on the original task, at least one of the three proposed strategies—Verbalized, CoT, and Decomposed—can effectively detect mismatches between visual and textual inputs. Notably, stronger models like GPT-4o achieve particularly high accuracy in conflict detection.

Our analysis indicates that the effectiveness of each mitigation strategy varies by both model and task. For instance, the decomposed method only appears effective when the model's performances using image-only and text-only inputs are reasonably high. Moreover, although both the Verbalized and CoT approaches excel when the model's performance on the original task is high, they exhibit distinct behavior: the Verbalized approach performs better for open models (despite overall low accuracy) and GPT-4o, whereas GPT-4o mini benefits more from the CoT method in the connectivity, convexity, and root calculation tasks. These differences highlight the nuanced nature of bias in VLM predictions and suggest that tailored mitigation techniques may be necessary to address specific challenges. Overall, our results underscore the need for a multifaceted approach to mitigate modality bias, ultimately enhancing the reliability and fairness of vision-language models.

Model	Connectivity		Convexity		Roots		Science		VSR	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
Qwen2 7B + V	49.6	7.1	53.5	36.7	50.3	14.6	58.5	51.7	57.0	68.0
Qwen2 7B + C	50.0	0.0	49.8	16.2	49.8	8.2	50.5	15.5	64.6	70.5
Qwen2 7B + D	-	-	-	-	-	-	-	-	-	-
Qwen2 72B + V	52.7	48.9	50.1	3.7	52.1	44.2	61.6	40.6	76.0	79.0
Qwen2 72B + C	48.8	25.9	50.9	27.2	52.7	33.6	54.5	19.6	80.5	79.0
Qwen2 72B + D	49.6	25.4	54.8	62.0	50.6	64.9	62.6	63.7	23.2	24.7
Llama-3.2 90B + V	50.7	41.6	56.4	60.1	55.3	62.8	68.1	58.2	45.8	35.1
Llama-3.2 90B + C	51.1	12.5	54.1	32.2	52.6	44.6	57.0	30.8	46.6	19.3
Llama-3.2 90B + D	52.7	52.5	60.7	62.9	53.5	63.6	68.1	71.2	10.2	9.9
GPT-4o mini + V	51.5	17.3	70.9	63.7	71.4	65.8	63.1	45.1	74.6	**79.2**
GPT-4o mini + C	56.3	39.8	82.8	82.2	**72.1**	63.1	61.6	40.6	54.6	55.6
GPT-4o mini + D	73.8	73.1	83.6	84.4	70.4	**77.1**	69.2	70.5	41.6	45.7
GPT-4o + V	67.6	55.6	**87.3**	**86.8**	72.0	62.7	78.3	75.7	**76.3**	77.2
GPT-4o + C	58.6	30.3	82.8	80.4	69.3	56.7	75.3	68.4	73.9	72.2
GPT-4o + D	**81.3**	**80.8**	85.0	85.9	68.3	75.9	**79.3**	**79.4**	68.7	68.7

Table 1: **The performance of the mitigation strategies** is evaluated by calculating the accuracy and F1 metrics for the three approaches, Verbalized (V), CoT (C), and Decomposed (D). We assess their effectiveness in reporting "mismatch" when a discrepancy exists between the input modalities.

5 Related Works

As LLMs and VLMs become more capable in reasoning tasks (Zhang et al., 2023; Ahn et al., 2024; Davoodi et al., 2024; Wang et al., 2024a), it is essential to investigate the interplay between modalities and how models reason across them to solve tasks. Previous studies have examined various aspects of VLM reasoning capabilities by introducing benchmarks that assess their general performance (Liu et al., 2024b; Yue et al., 2024), or by focusing on specific abilities such as spatial reasoning (Chen et al., 2024) and robot navigation (Zeng et al., 2023). Despite these efforts, it remains unclear how these models reason and combine information across multiple modalities, and whether they exhibit specific biases toward any input modality.

Many previous works investigate the robustness of VLMs (Chang et al., 2024). For instance, Yuksekgonul et al. (2022) demonstrate that VLMs are sensitive to the order of words, while Dumpala et al. (2024) show that they struggle to distinguish between semantically equivalent but lexically different captions. Furthermore, several studies have examined VLMs robustness through adversarial attacks (Liu et al., 2024a; Ye et al., 2025) and jailbreak strategies (Tao et al., 2024; Jin et al., 2024). The study most closely related to ours is a concurrent work (Deng et al., 2025) that examines VLMs bias toward text and image inputs in conflicting scenarios. Our approach differs in three key aspects. First, the authors in (Deng et al., 2025) generate mismatching text using GPT-4o; in contrast, we create mismatched text using rule-based and manual methods, giving us greater control. Second, while (Deng et al., 2025) report that VLMs exhibit a "blind faith in text", we demonstrate that VLMs can, in fact, show a stronger bias toward images in certain tasks, and the extent of their bias is heavily correlated with sample difficulty. Third, whereas their work focuses on supervised fine-tuning approaches for mitigation, which can be challenging and expensive, we explore the impact of post-processing mitigation strategies.

6 Conclusion

We investigate bias in VLMs by creating five novel benchmarks that span various tasks and domains, including mathematical reasoning, science, and visual description. Our findings reveal that VLMs exhibit distinct biases toward either textual or visual cues depending on the task, with the nature of this bias shifting according to sample difficulty and model capabilities—favoring text in simpler scenarios and images in more complex ones. Furthermore, we explore a range of mitigation strategies—including verbalized, chain-of-thought, and decomposed approaches—which have demonstrated reasonable accuracy in detecting mismatches, provided the model achieves strong performance on the original task. These results underscore the complexity of multimodal reasoning in VLMs and highlight promising directions for enhancing their reliability in real-world applications.

7 Limitations

Despite the insights provided by our study, several limitations must be acknowledged. First, our analysis is restricted to only five vision-language models. While these models represent a range of architectures and capabilities, they may not capture the full diversity of VLM behaviors present in the broader research community. Future work should expand this analysis to include additional models to validate and generalize our findings.

In addition, although our five benchmarks cover various tasks and domains—including mathematical reasoning, science, and visual descriptions—they may not fully represent the myriad of real-world scenarios where multimodal conflicts occur. Our benchmarks, while comprehensive in their scope, are still limited in terms of the diversity and complexity of tasks that can arise in practical applications.

Finally, while the mitigation strategies we investigated (Verbalized, CoT, and Decomposed) have shown promising results in some cases, they may not be universally applicable across all models and tasks. Further research is needed to refine these approaches and explore additional strategies that could enhance the robustness and fairness of vision-language models in handling conflicting information.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. *arXiv preprint arXiv:2402.00157*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. 2024. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465.

Arash Gholami Davoodi, Seyed Pouyan Mousavi Davoudi, and Pouya Pezeshkpour. 2024. Llms are not intelligent thinkers: Introducing mathematical topic tree benchmark for comprehensive evaluation of llms. *arXiv preprint arXiv:2406.05194*.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. 2025. Words or vision: Do vision-language models have blind faith in text? *arXiv preprint arXiv:2503.02199*.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Shama Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. 2024. Sugarcrepe++ dataset: Vision-language model sensitivity to semantic and lexical alterations. *Advances in Neural Information Processing Systems*, 37:17972–18018.

Deqing Fu, Ruohao Guo, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. 2024. Isobench: Benchmarking multimodal foundation models on isomorphic representations. *arXiv preprint arXiv:2404.01266*.

Alireza Ghafarollahi and Markus J Buehler. 2024. Atomagents: Alloy design and discovery through physics-aware multi-modal multi-agent artificial intelligence. *arXiv preprint arXiv:2407.10022*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Bowen Jiang, Yangxinyu Xie, Xiaomeng Wang, Weijie J Su, Camillo Jose Taylor, and Tanwi Mallick. 2024. Multi-modal and multi-agent systems meet rationality: A survey. In *ICML 2024 Workshop on LLMs and Cognition*.

Haibo Jin, Leyang Hu, Xinuo Li, Peiyan Zhang, Chonghan Chen, Jun Zhuang, and Haohan Wang. 2024. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv preprint arXiv:2407.01599*.

Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. 2025. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*.

Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. 2024a. A survey of attacks on large vision-language models: Resources, advances, and future trends. *arXiv preprint arXiv:2407.07403*.

Fangyu Liu, Guy Emerson, and Nigel Collier. 2023a. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024b. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Xijia Tao, Shuai Zhong, Lei Li, Qi Liu, and Lingpeng Kong. 2024. Imgtrojan: Jailbreaking vision-language models with one image. *arXiv preprint arXiv:2403.02910*.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024a. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. 2025. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and 1 others. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Jianhao Yuan, Shuyang Sun, Daniel Omeiza, Bo Zhao, Paul Newman, Lars Kunze, and Matthew Gadd. 2024. Rag-driver: Generalisable driving explanations with retrieval-augmented in-context learning in multi-modal large language model. *arXiv preprint arXiv:2402.10828*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 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.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Fanlong Zeng, Wensheng Gan, Yongheng Wang, Ning Liu, and Philip S Yu. 2023. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*.

Task	# Samples
Connectivity	128
Convexity	256
Roots	400
Science	99
VSR	846

Table 2: Data statistics of the created benchmarks.

Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348.

A Data

We build our dataset on top of two existing sources. The first, VSR, contains over 10k natural text (one-sentence) and image pairs, annotated with 66 types of spatial relations and true/false labels indicating whether the text and image are aligned. The second, Isobench, is a dataset featuring problems from math, science, algorithms, and games, with each example presented in multiple isomorphic formats (e.g., visual and textual). The data statistics for our created datasets are provided in Table 2.

B Prompts

The prompt used to extend the VSR description is provided in B.1. Additionally, the prompt for solving tasks using both image and text inputs is presented in B.2, B.3, B.4, B.5 and B.6. We also include the prompts for solving the graph connectivity task with single modality inputs in B.7 and B.8. Moreover, the prompts employed in the mitigation strategies for graph connectivity are provided in B.9, B.10, B.11, B.12, and B.13.

VSR Description Extension

```
Below is a sentence that serves as the central idea, along with a specified relationship between entities within that sentence. Given an image, your task is to create a complete, coherent paragraph by adding few sentences before and after the given sentence describing the image. The additional sentences should establish context in a logical manner. However, these extra sentences should not directly connect to the provided relationship. Ensure that the specific relationship appears exactly once as given within the paragraph. Finally, do not mention 'image' in the paragraph.

Given sentence:
{}
Specified relationship:
{}
```

Connectivity - Pair Input

```
You are given the adjacency matrix of a graph and two target nodes. The goal is to determine whether the target nodes are connected. Additionally, you are provided with an image of the graph, where the target nodes are highlighted in {}. Use both the image and the adjacency matrix to answer the question.
The output should be in the following format:
Explanation: <....>
Answer: <True or False>

Adjacency matrix:
{}

Target nodes = {} and {}
```

Convexity - Pair Input

```
You are given the mathematical expression of a function. The goal is to determine whether the function is convex or concave. Additionally, you are provided with the plot of the function. Use both the plot and the mathematical expression of the function to identify the convexity of the function.
The output should be in the following format:
Explanation: <....>
Answer: <convex or concave>

The mathematical expression:
{}
```

Roots - Pair Input

```
You are given the mathematical expression of a polynomial. The goal is to find the roots of this polynomial. Additionally, you are provided with an image of the plot of that polynomial, where the roots are highlighted in red. Use both the image and the the mathematical expression of a polynomial to find the roots.
The output should be in the following format:
Explanation: <....>
Answer: <[The list of roots, comma separated]>

The polynomial:
{}
```

Science - Pair Input

```
You are given a question, multiple choices, a textual description, and an image. Your task is to analyze both the text and the image and then choose the answer of the question from the provided choices.
The output should be in the following format:
Explanation: <....>
Answer: <A or B or ...>

The textual description:
{}

The question:
{}

The Choices:
{}
```

VSR - Pair Input

```
You are given a statement, a paragraph, and an image. Your task is to analyze both the paragraph and the image to decide whether the statement is true or false.
The output should be in the following format:
Explanation: <....>
Answer: <true or false>

The paragraph:
{}

The statement:
{}
```

Connectivity - Only Text

```
You are given the adjacency matrix of a graph and two target nodes. The goal is to determine whether the target nodes are connected.
The output should be in the following format:
Explanation: <....>
Answer: <True or False>

Adjacency matrix:
{}

Target nodes = {} and {}
```

Connectivity - Only Image

```
You are provided with an image of a graph, with two target nodes highlighted in {}. The goal is to determine whether the target nodes are connected.
The output should be in the following format:
Explanation: <....>
Answer: <True or False>

Target nodes = {} and {}
```

Mitigation-V - Connectivity

```
You are given the adjacency matrix of a graph and two target nodes. The goal is to determine whether the target nodes are connected. Additionally, you are provided with an image of the graph, where the target nodes are highlighted in {}. Use both the image and the adjacency matrix to answer the question.
If there is a mismatch or a contradiction between image and text, the answer should be 'Mismatch'.
The output should be in the following format:
Explanation: <....>
Answer: <True or False or Mismatch>

Adjacency matrix:
{}
Target nodes = {} and {}
```

Mitigation-C - Connectivity

```
You are given the adjacency matrix of a graph and two target nodes. The goal is to determine whether the target nodes are connected. Additionally, you are provided with an image of the graph, where the target nodes are highlighted in {}. Use both the image and the adjacency matrix to answer the question.

Follow these steps to find the answer:

Step 1: Attempt to determine the answer using only the image. If the image alone does not provide the answer, extract all relevant information from it.
Step 2: Attempt to determine the answer using only the text. If the text alone does not provide the answer, extract all relevant information from it.
Step 3: Compare the answers or extracted information from both the text and the image, combine them, and derive the final answer.

If there is a mismatch or a contradiction between image and text, the answer should be 'Mismatch'.
The output should be in the following format:
Explanation: <....>
Answer: <True or False or Mismatch>

Adjacency matrix:
{}
Target nodes = {} and {}
```

Mitigation-D (Combining Module) - Connectivity

```
Given the adjacency matrix of a graph, two target nodes, and the image of the graph, the goal was to determine whether the target nodes are connected.
We provide only the task's textual description to an LLM to obtain an answer, and separately supply only the visual description to a VLM to obtain its answer.
Your goal is to compare the answers from the both models, combine them, and derive the final answer.
If there is a mismatch or a contradiction between two answers, the final answer should be 'Mismatch'.
The output should be in the following format:
Explanation: <....>
Answer: <True or False or Mismatch>

LLM's Answer:
{}

VLM's Answer:
{}
```

Mitigation-D (Image Module) - Connectivity

```
You are provided with an image of a graph, with two target nodes highlighted in {}. The goal is to determine whether the target nodes are connected.
If the provided information is insufficient to solve the task, extract all relevant details from the input.
The output should be in the following format:
Explanation: <....>
Answer: <True or False>

Target nodes = {} and {}
```

Mitigation-D (Text Module) - Connectivity

```
You are given the adjacency matrix of a graph and two target nodes. The goal is to determine whether the target nodes are connected.
If the provided information is insufficient to solve the task, extract all relevant details from the input.
The output should be in the following format:
Explanation: <....>
Answer: <True or False>

Adjacency matrix:
{}

Target nodes = {} and {}
```

C VLMs Performance on The datasets

We provide the accuracy of VLMs on each task in Table 3. For the root calculation task, we report the average accuracy. First, the performance on both the original and modified text is nearly identical across all models, highlighting the minimal modifi-

cations applied to create conflicting textual cues. In most cases, models perform better using textual representations alone compared to using images alone, which aligns with observations from the original papers. However, for connectivity, GPT-4o, for convexity, Qwen2 72B, and for root calculation, both GPT-4o and GPT-4o mini, the image input appears to be more helpful than text alone. Finally, Qwen2 7B was unable to solve the task when the image was missing.

D VLMs Bias Value

We calculate the bias value for VLMs as the difference between the percentage of image-favored responses and the percentage of text-favored responses, with results provided in Table 4. Positive values indicate a preference for images, while negative values reflect a preference for textual input. Our findings reveal significant variation across models and tasks, with bias values ranging from 56.8% in favor of images to 74.4% in favor of text. In general, less capable models tend to exhibit a stronger bias toward textual data. Moreover, VLMs show the highest bias toward text in the VSR dataset, whereas they exhibit a greater degree of bias toward images in the graph connectivity task.

E Error Analysis

We present a Venn diagram illustrating GPT-4o mini, Llama-3.2 90B's, Qwen2 72B's, Qand wen2 7B's error percentages under three different input settings: text-only, image-only, and combined matching initial text-image inputs in Figures 7, 8, 9, and 10, respectively. We observe similar patterns to those seen with GPT-4o, particularly regarding the correlation between the models' unique and common error counts and their reported bias across different tasks.

	Model	Only Text Original	Only Text Modified	Only Image	Pair Text-Image Original
Connectivity	Qwen2 7B	-	-	48.4	53.1
	Qwen2 72B	50.7	54.6	50.7	53.9
	Llama-3.2 90B	57.8	58.5	55.4	65.6
	GPT-4o mini	86.7	80.4	84.3	76.5
	GPT-4o	92.9	84.3	96.0	96.0
Convexity	Qwen2 7B	-	-	19.5	71.8
	Qwen2 72B	61.7	60.5	73.4	87.5
	Llama-3.2 90B	85.9	86.3	54.2	90.6
	GPT-4o mini	95.3	95.3	91.4	98.0
	GPT-4o	94.1	90.2	92.5	97.2
Roots	Qwen2 7B	-	-	31.2	36.1
	Qwen2 72B	28.2	29.6	24.6	33.8
	Llama-3.2 90B	31.6	29.3	16.7	35.6
	GPT-4o mini	38.6	37.4	96.8	97.5
	GPT-4o	34.7	34.3	98.8	99.2
Science	Qwen2 7B	-	-	60.6	84.8
	Qwen2 72B	89.8	86.8	54.5	89.8
	Llama-3.2 90B	87.8	84.8	50.5	89.8
	GPT-4o mini	87.8	82.8	67.6	87.8
	GPT-4o	92.9	89.8	84.8	91.9
VSR	Qwen2 7B	-	-	7.8	79.3
	Qwen2 72B	97.5	97.9	30.6	79.6
	Llama-3.2 90B	93.1	98.3	12.6	79.3
	GPT-4o mini	96.6	99.1	35.9	70.5
	GPT-4o	98.8	95.9	67.7	86.1

Table 3: The accuracy of models in solving each task when provided with text and image inputs independently and in combination.

Model	Connectivity	Convexity	Roots	Science	VSR
Qwen2 7B	37.5	-7.8	-44.4	6.0	-71.9
Qwen2 72B	38.2	-39.8	-29.0	-44.4	-74.4
Llama-3.2 90B	-7.8	-53.9	-29.4	-46.4	-71.9
GPT-4o mini	23.4	-53.5	56.8	-42.4	-54.9
GPT-4o	52.3	-65.2	42.3	19.1	6.9

Table 4: **VLMs' bias value** for each dataset, calculated as the difference between the percentage of image-favored responses and the percentage of text-favored responses. Positive values indicate an image preference, while negative values indicate a text preference.

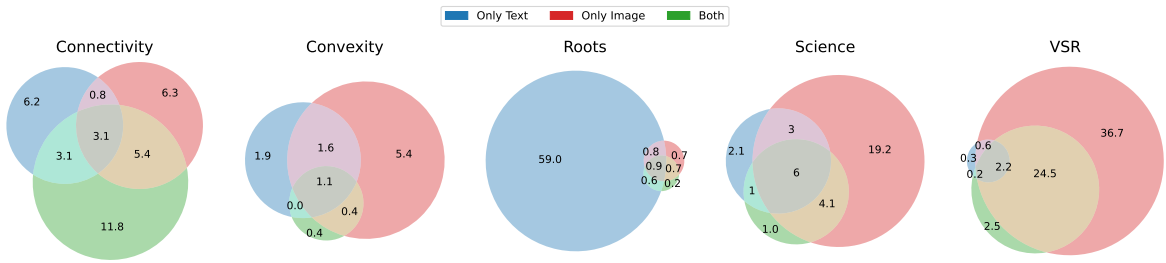


Figure 7: **Venn diagram of errors** displaying the percentage of mispredicted samples when GPT-4o mini is provided with only text, only image, and both text and image inputs.

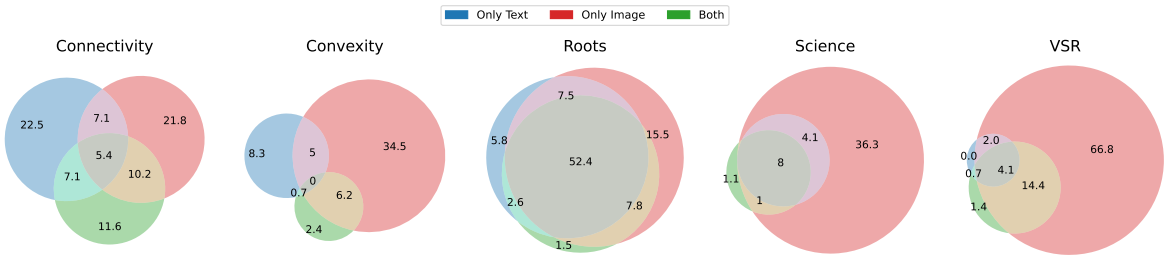


Figure 8: **Venn diagram of errors** displaying the percentage of mispredicted samples when Llama-3.2 90B is provided with only text, only image, and both text and image inputs.

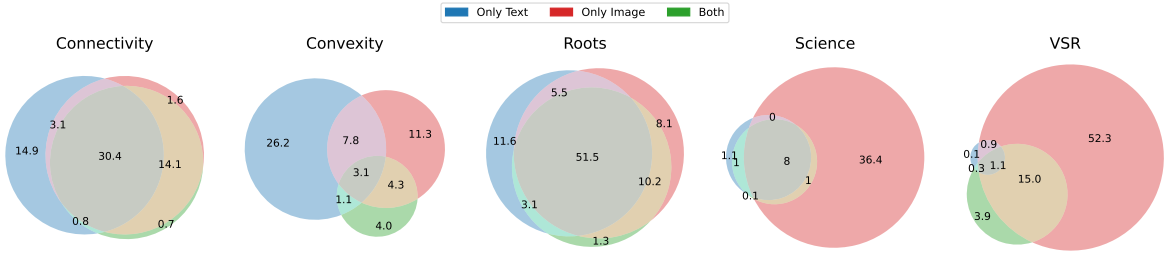


Figure 9: **Venn diagram of errors** displaying the percentage of mispredicted samples when Qwen2 72B is provided with only text, only image, and both text and image inputs.

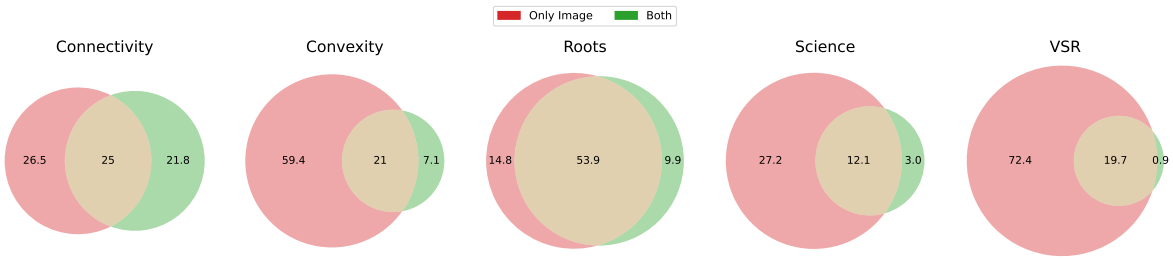


Figure 10: **Venn diagram of errors** displaying the percentage of mispredicted samples when Qwen2 7B is provided with only text, only image, and both text and image inputs.