On the Effect of Physical Integration of Textual and Visual Information on Reasoning in Multi-modal Large Language Models

Anonymous ACL submission

Abstract

We introduce SPLINT, a novel diagnostic benchmark designed to investigate whether physical integration of textual and visual information improves reasoning in MLLMs. Motivated by cognitive load theory, SPLINT consists of 925 paired visual question-answer instances under both integrated and split instruction settings. Using this benchmark, we evaluate the reasoning capabilities of 12 popular MLLMs across the two settings. Our experiments reveal that physical integration of textual and visual information does not consistently enhance reasoning performance and can even degrade performance in certain cases. Further qualitative analysis shows that integrated instructions may interfere with precise visual grounding, leading MLLMs to misinterpret visual elements and propagate these perceptual errors throughout the reasoning process. These findings suggest that physical integration is not universally beneficial for MLLMs and should therefore be applied with careful consideration depending on the model and task domain. We hope that SPLINT will serve as a valuable benchmark for future research on understanding and improving multi-modal reasoning in MLLMs.[1]

1 Introduction

Multi-modal large language models (MLLMs) have shown strong performance across a wide range of visual reasoning tasks (Yin et al., 2024; Zhang et al., 2024a). Consequently, increasing attention has been devoted to understanding how MLLMs reason under diverse practical settings involving different input modalities (Li et al., 2025b). Among these settings, recent studies have increasingly focused on the vision-only setting (Yue et al., 2025; Miyai et al., 2025), where both textual and visual information are presented within an image. Notably, this setting reflects realistic scenarios in which users provide instructions in a unified visual format, such as providing scientific figures or instruction manuals formatted as screenshots, photographs, or scanned documents (Li et al., 2024b,d; Zhang et al., 2023; Long et al., 2025; Li et al., 2026).

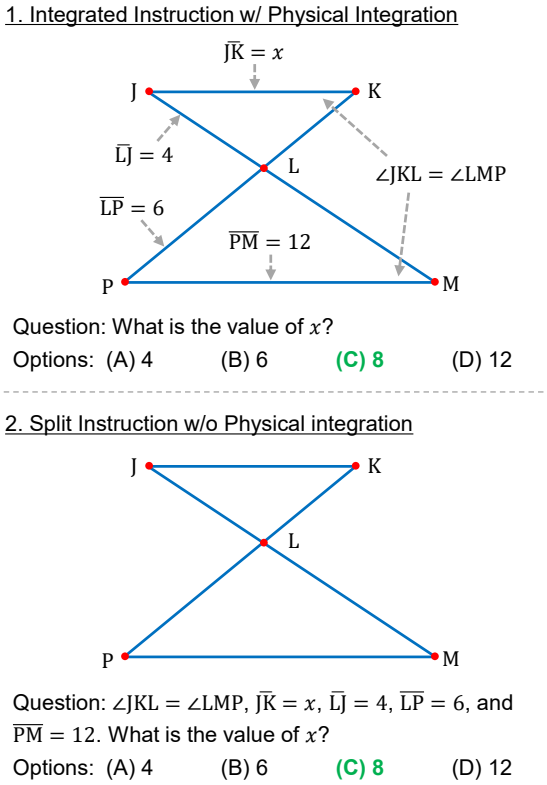


Figure 1: Comparison of (a) integrated and (b) split instructions for a geometry problem. The integrated instruction involves physical integration of textual and visual information, whereas the split instruction does not. Green text indicates the correct answer. See Appendix A for additional examples from other domains.

However, the spatial arrangement of instructions in this setting is predominantly designed to facilitate human reasoning (Sweller et al., 1998), and it remains unclear whether such arrangements are equally effective for MLLM reasoning. If a strategy

[1] Our code is available at https://anonymous.4open.science/r/SPLINT-3BDB/README.md.

for arranging instructions originally intended to improve human reasoning instead impairs MLLM reasoning, applying the same strategy to instructions for MLLMs may not be an optimal choice, and it may be beneficial to rather discard the strategy for MLLMs. To resolve this uncertainty and identify instruction arrangements that better support MLLM reasoning, it is important to first investigate how existing strategies for arranging instructions affect reasoning in MLLMs.

In cognitive load theory, one prominent strategy for arranging instructions to facilitate human reasoning is physical integration of textual and visual information (Chandler and Sweller, 1991; Sweller et al., 1990). As shown in Figure 1, physical integration can be achieved by placing textual explanations (e.g., descriptions of the side lengths or angles of a triangle) close to the relevant visual elements (e.g., line segments in a triangle) and/or by explicitly linking them using visual artifacts such as arrows. Focusing on this well-established strategy, we ask the following underexplored research question: under vision-only settings, how does physical integration of textual and visual information affect MLLM reasoning?

To address this research question, we introduce SPLINT (**Spl**it and **Int**egrate), a novel benchmark designed to analyze how physical integration of textual and visual information affects reasoning in MLLMs. SPLINT consists of 925 pairs of visual reasoning QA instances. In each pair, one instance presents an integrated instruction, where textual and visual information are physically integrated, while the other presents a split instruction, where such integration is not applied (see Figure 1 for examples). By constructing the benchmark as paired QA instances, we enable controlled comparisons of MLLM reasoning with and without physical integration.

Using SPLINT, we evaluate 12 popular MLLMs under both integrated and split instruction settings. Our experimental results show that physical integration of textual and visual information does not consistently improve performance and can even degrade it in some cases. Further analysis reveals that MLLMs often suffer from perceptual errors under the integrated instruction setting, misinterpreting visual elements and subsequently propagating these errors through the reasoning process. These findings contrast with cognitive science research showing the benefits of physical integration for human reasoning. Taken together, our results suggest that physical integration is not universally beneficial for MLLMs and should therefore be applied with careful consideration depending on the model and task domain.

Our contributions are summarized as follows:

- We are the first to explore the effect of physical integration of textual and visual information on MLLM reasoning, drawing inspiration from cognitive science research.
- We propose SPLINT, a novel benchmark designed to analyze how physical integration affects reasoning in MLLMs.
- We conduct experiments on recent MLLMs and show that physical integration of textual and visual information does not consistently improve reasoning performance.

2 Related Works

2.1 Multi-modal Instruction Design

Prior work has explored various multi-modal instruction designs to understand their effects on MLLM reasoning. In particular, a line of research improves MLLM reasoning by providing instructions that encourage models to perform subtasks beneficial for solving the original task. These subtasks include scene graph generation (Mitra et al., 2024), sketch generation (Hu et al., 2024; Xiao et al., 2024; Wang et al., 2025; Wu et al., 2024; Izadi et al., 2025; Vinker et al., 2025), subquestion answering (Zheng et al., 2023; You et al., 2023; Zhou et al., 2024; Jiang et al., 2024b; Li et al., 2024e; Acuna et al., 2025), visual token selection (Gao et al., 2025; Liu et al., 2026), and reasoning structure discovery (Wu et al., 2026). While these studies investigate how different subtasks affect reasoning performance, our work takes a different direction by focusing on how the arrangement of textual and visual information within an image influences MLLM reasoning, rather than introducing specific subtasks.

Another line of work (Yue et al., 2025; Miyai et al., 2025; Li et al., 2024c) studies the vision-only setting, where textual information such as questions and answer choices is embedded within an image rather than provided to models as separate modality inputs. These works reveal weaknesses of current MLLMs by showing that they often struggle to reason in this setting compared to non-vision-only

settings. However, they largely overlook how different arrangements of textual and visual information within an image affect reasoning performance in the vision-only setting. In contrast, we focus on physical integration of textual and visual information as a particular arrangement strategy inspired by cognitive science research and investigate how this strategy influences MLLM reasoning.

2.2 Benchmarks for MLLMs

Prior work has proposed numerous benchmarks to evaluate the visual reasoning capabilities of MLLMs. Existing benchmarks can be broadly categorized into two groups based on the level and type of reasoning they emphasize. The first group focuses on general-purpose visual reasoning abilities, including visual perception and commonsense reasoning. Representative benchmarks include LAMM (Yin et al., 2023), SEED-Bench (Li et al., 2024a), LVLM-eHub (Xu et al., 2025), MMBench (Liu et al., 2024b), CV-Bench (Tong et al., 2024), MM-Vet (Yu et al., 2024), Mantis (Jiang et al., 2024a), BLINK (Fu et al., 2024), and Visual Riddles (Guetta et al., 2024).

In contrast, the second group emphasizes more challenging reasoning tasks that require expert-level or discipline-specific knowledge. For example, MMMU (Yue et al., 2024) and M3CoT (Chen et al., 2024) evaluate multi-disciplinary visual reasoning abilities across diverse domains. MMMU-Pro (Yue et al., 2025) further increases the reasoning difficulty by filtering out questions that can be answered solely from textual information, while MMReason (Yao et al., 2025) applies multi-model voting techniques to reduce shortcut solutions based on guessing or memorization. Other benchmarks investigate specific aspects of reasoning, such as CoT reasoning (Jiang et al., 2025) and reasoning that requires generating multi-modal outputs (Guo et al., 2025). In addition, several benchmarks focus specifically on STEM reasoning tasks (Lu et al., 2024; He et al., 2024; Li et al., 2025a; Xiang et al., 2025) to assess specialized reasoning abilities.

Despite these advances, existing benchmarks primarily emphasize task difficulty and domain coverage, while paying limited attention to how the arrangement of textual and visual information within an image affects reasoning performance. Unlike these benchmarks, our work systematically investigates how physical integration of textual and visual information influences MLLM reasoning.

3 Background and Motivation

In this section, we first introduce the concept of physical integration in cognitive load theory (Section 3.1) and then explain our motivations for adopting it for MLLM evaluation (Section 3.2).

3.1 Background: Physical Integration in Cognitive Load Theory

Cognitive load theory suggests that effective arrangements of textual and visual information improve human understanding by reducing cognitive load (Chandler and Sweller, 1991; Sweller et al., 1990). One notable strategy for such arrangements is physical integration, which involves placing textual explanations close to the relevant visual elements and/or explicitly linking them using visual cues such as arrows. For humans, this form of integration can reduce the need for mental integration across disparate sources of information, thereby improving comprehension compared to split instructions in which textual and visual information are not physically integrated.

3.2 Motivation for Introducing Physical Integration to MLLM Evaluation

As noted in the previous subsection, physical integration can reduce the burden of mental integration for humans when they read instructions. However, it is vital to clarify that we do not aim to investigate whether MLLMs can simulate the human cognitive processes underlying this reduction. More fundamentally, we do not aim to define or implement mental integration in MLLMs. Instead, our goal is to study how physical integration influences reasoning outputs generated by MLLMs.

Exploring the effect of physical integration on MLLM reasoning is important for enhanced human-MLLM collaboration. Due to its effectiveness in improving human understanding of instructions, physical integration has been widely used to arrange instructions in materials such as annotated scientific figures (Li et al., 2024b,d) and instruction manuals (Zhang et al., 2023; Long et al., 2025; Li et al., 2026). In addition, MLLMs have increasingly been used in collaboration with humans to solve problems (Durante et al., 2024; Zheng et al., 2025; Liu et al., 2025). Thus, MLLMs may frequently encounter instructions with physical integration during such collaboration. Given these trends, investigating the impact of physical integration on MLLM reasoning could help iden-

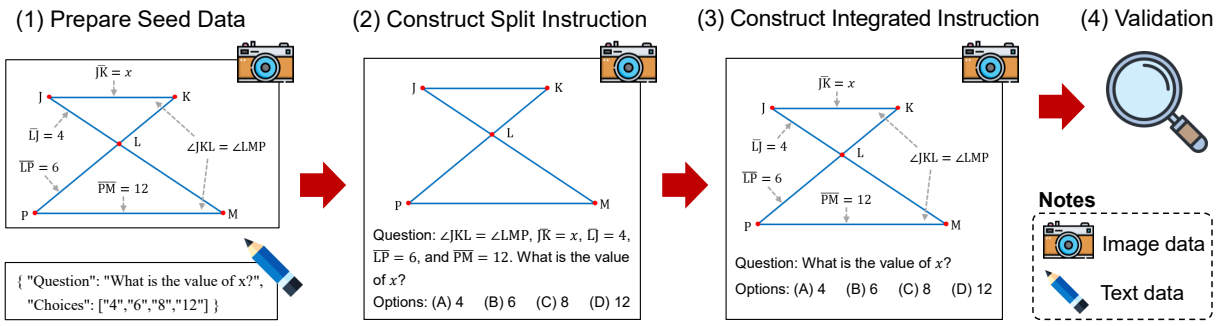


Figure 2: Data collection pipeline for SPLINT. (1) Given seed data consisting of an image, a question, and answer options, (2) textual explanations are extracted from the image and consolidated into the question to create split instructions. (3) These statements are then repositioned near their corresponding visual elements and/or explicitly linked to those visual elements using artifacts such as arrows, resulting in integrated instructions. (4) Lastly, constructed instructions are validated by the authors to ensure data quality.

tify instruction arrangements that facilitate better MLLM reasoning, thereby supporting more effective human-MLLM collaboration.

4 SPLINT Construction

In this section, we describe the data collection process for constructing SPLINT. As illustrated in Figure 2, the pipeline consists of four stages: (1) preparing seed data, (2) constructing split instructions, (3) constructing integrated instructions, and (4) validation.

4.1 Prepare Seed Data

We collect seed data from existing benchmarks in geometry, mechanical engineering, and electrical engineering. These three domains are selected because they are among the domains in which physical integration of textual and visual information has been shown to improve human reasoning (Chandler and Sweller, 1991). Specifically, we use Geometry3K (Lu et al., 2021) for geometry, the mechanical engineering subset of MMMU (Yue et al., 2024) for mechanical engineering, and EEE-Bench (Li et al., 2025a) for electrical engineering. Each sample in these benchmarks consists of an image, a question, and a set of answer options.

4.2 Construct Split Instructions

In this stage, we construct split instructions where textual explanations remain separated from the diagram. Given a sample consisting of an image, a question, and answer options, we extract textual explanations embedded within the image and relocate them into the question text. For geometry samples, we utilize the diagram formal language annotations provided by the dataset to automate

Domains	Number of Instruction Pairs	Avg. Number of Textual Explanations
Geometry	511	4.54
Mechanical Eng.	162	4.98
Electrical Eng.	252	4.60
Total	925	4.64

Table 1: Basic statistics of the SPLINT benchmark.

this process. For mechanical and electrical engineering samples, we manually inspect each image and extract the relevant textual explanations. As a result, textual information that was originally intertwined with visual elements is consolidated into a separate textual region, resulting in a separation between textual and visual information.

4.3 Construct Integrated Instructions

Next, we construct integrated instructions, where textual explanations are explicitly associated with their corresponding visual elements through visual linkage. Starting from the split instructions, we reposition textual explanations so that they are placed near the relevant visual components and/or explicitly connected to those visual elements using artifacts such as arrows. For geometry samples, we leverage the annotated coordinates of points provided by the dataset to automatically determine appropriate placements for the textual explanations. For mechanical and electrical engineering samples, we manually inspect each image, following the procedure used in the previous stage.

4.4 Validation

To ensure data quality, the first author manually validates each constructed instruction. Specifically, we verify that the correct answers for both the inte-

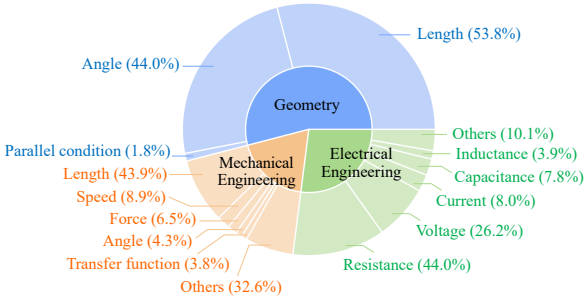


Figure 3: Distribution of the types of textual explanations that are relocated for physical integration. Percentages are normalized within each domain such that the total sums to 100%. For each domain, at most five of the most frequent types are shown.

grated and split instructions remain consistent with the original QA sample. We exclude instruction pairs if the answer to either instruction changes due to ambiguities or noise in the original QA samples. To maintain a controlled evaluation environment, we standardize the image resolution, diagram size, and text size for each pair of split and integrated instructions. When necessary, additional white space padding is added to unify image resolutions across samples. Table 1 presents the basic statistics of the SPLINT benchmark.

4.5 Dataset Analysis

To better understand the domain-specific characteristics covered in SPLINT, we conduct a dataset analysis by examining the distributions of textual explanation types and answer types across domains. Figure 3 shows the distribution of textual explanation types that are relocated for physical integration. We observe that SPLINT contains diverse and domain-specific explanations across domains. For instance, geometry problems frequently involve descriptions of geometric properties such as angles and lengths. The engineering domains exhibit even broader varieties of explanations, including concepts such as lengths (covering distance-related properties such as width, depth, and height), speed, force, resistance, and voltage.

A similar tendency can also be observed in Figure 9, which presents the distributions of answer types in SPLINT. The answer distributions differ substantially across domains and involve a wide range of terminologies. These results suggest that SPLINT covers a broad range of domain-specific concepts, requiring MLLMs to accurately interpret heterogeneous textual explanations grounded in visual information.

5 Experiments

5.1 Experimental Settings

Baseline MLLMs: In the experiment, we employ 7 recent MLLMs to establish our baselines. For open-source MLLMs, we use Qwen2.5-VL (7B, 32B, 72B) (Bai et al., 2025), Qwen3-VL (8B, 32B) (Team, 2025b), Gemma3 (12B, 27B) (Team, 2025a), and LLaVA-NeXT (13B, 72B) (Liu et al., 2024a). For closed-source MLLMs, we experiment with GPT-4o, GPT-4o-mini (OpenAI, 2024), and GPT-5-mini (OpenAI, 2025). We perform inference with the open-source models using Hugging Face and access the closed-source models through the OpenAI API.[2] We set the temperature of $\tau = 0.0$ and top-p of 1.0 for the closed-source models.

Prompting Methods: We adopt two widely used prompting methods: a direct and Chain-of-Thought (CoT) prompting (Kojima et al., 2022). The direct prompting asks the model to select the correct option without any intermediate reasoning, whereas the CoT prompting encourages step-by-step reasoning by appending "Think step by step." Following Yue et al. (2025), we report the higher ones in the main results. See Table 3 for detailed prompt designs.

Evaluation Metrics: To enable a clean and consistent evaluation, all tasks in SPLINT are formulated as multiple-choice question answering. Accordingly, we adopt standard accuracy as the evaluation metric.

5.2 Main Results

Table 2 presents the performance of MLLMs on SPLINT. For each result, we report the mean performance and standard deviation across five runs. Some MLLMs in LLaVA family can suffer from repetitive generation, which can reduce their performance below 25% random accuracy. Overall, we observe statistically significant performance differences between the two instruction types in a majority of cases, although the sign and magnitude of these differences vary considerably across models. For example, GPT-4o-mini achieves higher accuracy under the integrated instruction than under the split instruction, with improvements ranging from +3.2 to +3.8%. In contrast, Qwen2.5-VL-7B exhibits performance degradation under the integrated instruction in two of the three domains, with

[2] https://platform.openai.com/docs/api-reference/

Model	Geometry			Mechanical Engineering			Electrical Engineering		
	Int.	Split	Δ	Int.	Split	Δ	Int.	Split	Δ
Open-source LLMs									
Qwen2.5-VL-7B (2025)	$57.8_{\pm1.1}$	$60.5_{\pm1.1}$	-2.7*	$51.9_{\pm1.1}$	$53.7_{\pm0.9}$	-1.8*	$30.5_{\pm1.7}$	$33.4_{\pm1.1}$	-3.0*
Qwen3-VL-8B (2025b)	$74.1_{\pm1.4}$	$76.3_{\pm1.2}$	-2.2*	$42.4_{\pm1.9}$	$45.3_{\pm2.0}$	-2.9*	$39.1_{\pm1.2}$	$38.3_{\pm1.3}$	$+0.8$
Gemma3-12B (2025a)	$68.8_{\pm2.5}$	$65.2_{\pm1.9}$	$+3.4$*	$40.0_{\pm3.3}$	$39.6_{\pm1.6}$	$+0.4$	$34.8_{\pm0.9}$	$29.8_{\pm2.9}$	$+5.0$*
LLaVA-NeXT-13B (2024a)	$13.9_{\pm1.4}$	$14.9_{\pm1.5}$	-1.0	$17.9_{\pm0.7}$	$21.2_{\pm1.1}$	-3.3*	$14.3_{\pm0.8}$	$15.0_{\pm1.0}$	-0.7
Gemma3-27B (2025a)	$73.4_{\pm2.3}$	$73.7_{\pm2.1}$	-0.3	$43.7_{\pm2.9}$	$49.8_{\pm4.1}$	-6.1*	$37.6_{\pm2.3}$	$41.4_{\pm1.2}$	-3.8*
Qwen2.5-VL-32B (2025)	$69.0_{\pm0.6}$	$72.2_{\pm0.9}$	-3.2*	$56.8_{\pm1.2}$	$60.4_{\pm2.5}$	-3.6*	$45.1_{\pm1.0}$	$44.4_{\pm0.5}$	$+0.7$
Qwen3-VL-32B (2025b)	$83.6_{\pm0.7}$	$83.1_{\pm0.3}$	$+0.5$	$41.0_{\pm2.3}$	$44.8_{\pm2.2}$	-3.8*	$48.1_{\pm1.6}$	$50.2_{\pm0.9}$	-2.1*
Qwen2.5-VL-72B (2025)	$72.9_{\pm0.7}$	$72.8_{\pm0.8}$	$+0.1$	$59.4_{\pm0.8}$	$60.6_{\pm0.7}$	-1.2*	$39.1_{\pm1.0}$	$41.6_{\pm1.4}$	-2.5*
LLaVA-NeXT-72B (2024a)	$27.0_{\pm2.0}$	$27.0_{\pm1.6}$	$+0.0$	$29.2_{\pm0.5}$	$32.7_{\pm1.4}$	-3.5*	$25.9_{\pm0.7}$	$30.0_{\pm0.9}$	-4.1*
Closed-source LLMs									
GPT-4o-mini (2024)	$55.4_{\pm1.3}$	$52.2_{\pm0.7}$	$+3.2$*	$47.6_{\pm2.1}$	$43.8_{\pm2.0}$	$+3.8$*	$38.2_{\pm3.1}$	$36.6_{\pm1.3}$	$+1.6$
GPT-4o (2024)	$62.7_{\pm1.8}$	$62.1_{\pm1.1}$	$+0.6$	$49.3_{\pm2.2}$	$48.6_{\pm3.7}$	$+0.7$	$45.1_{\pm1.1}$	$41.9_{\pm1.2}$	$+3.2$*
GPT-5-mini (2025)	$90.4_{\pm0.8}$	$90.7_{\pm0.6}$	-0.3	$50.0_{\pm0.6}$	$49.6_{\pm1.5}$	$+0.4$	$73.8_{\pm0.6}$	$75.8_{\pm0.6}$	-2.0*

Table 2: Performance of MLLMs on SPLINT. Standard deviations are reported as subscripts. "Int." denotes integrated instructions, and Δ represents the performance difference between integrated and split instructions. * indicates a statistically significant difference ($p < 0.05$) based on a t-test.

declines ranging from -1.8 to -3.0%.

One possible explanation for these results is that physical integration provides explicit guidance linking textual explanations to their corresponding visual elements, thereby reducing the burden of visual grounding—the process of aligning textual descriptions with relevant image regions. However, leveraging such explicit guidance also requires models to interpret visual artifacts such as arrows and spatial cues. If models are not sufficiently trained to utilize these signals, integration may instead introduce additional complexity, making it more difficult to identify correspondences between textual explanations and visual elements.

The results indicate that physical integration of textual and visual information does not consistently lead to performance improvements across models and domains. In several cases, applying physical integration to the instructions even resulted in performance degradation. Although the semantic content of the problems remains unchanged between the two instruction formats, many MLLMs failed to maintain robust reasoning performance. These findings suggest that physical integration is not universally beneficial for MLLMs and should therefore be applied with careful consideration depending on the model and task domain.

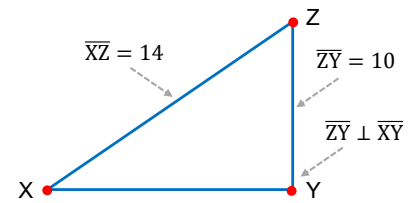
5.3 Error Analysis

To gain deeper insight into the reasoning errors of MLLMs on SPLINT, we conduct an error analysis of MLLMs under both integrated and split instructions, analyzing the CoT reasoning trajectories of Qwen2.5-VL-32B. Specifically, we first present the distribution of MLLM errors on integrated instructions in Figure 6. Through inspection on these error cases, we identify two major types of errors: 1) Geometric/Physical Property Misinterpretation, and 2) Incorrect Equation Derivation.

In the former type of error, the model misinterprets visual elements and subsequently propagates these errors through the reasoning process. Examples are shown in Figure 4. In the geometry example, the model incorrectly identifies the side $\overline{ZY}$ as the side opposite to angle $\angle Z$, although it is in fact the adjacent side. This perceptual mistake leads the model to apply an incorrect trigonometric relation and produce an erroneous final answer. Similarly, in the mechanical engineering example, the model confuses the diameter of the tank with its depth, resulting in incorrect volume estimation. These errors can be interpreted as perceptual errors (Yue et al., 2024), where the model incorrectly understand what is shown in the image for reasoning.

In the latter type of error, on the other hand, the model successfully interpret key visual elements but establish the wrong or irrelevant equation, which eventually results in incorrect final prediction. An Example is shown in Figure 5. In this electrical engineering example, the model correctly interprets all the voltage values, but incorrectly formulate the equation by directly computing the difference between two voltage values. These errors can be interpreted as reasoning errors (Yue et al., 2024), where the model understand the visual ele-

1. Geometry (Geometric/Physical Property Misinterpretation)



Question:
Find the measure of $\angle Z$ to the nearest tenth.

Options:
(A) 35.5 (B) 44.4 (C) 45.6 (D) 54.5

Ground Truth: (B)

Qwen2.5-VL-32B (Integrated):
1. **Identify the given information:**
- $\overline{XZ}$ = 14 (hypotenuse), $\overline{ZY}$ = 10 (opposite side to $\angle Z$)
- $\overline{ZY} \perp \overline{XY}$ (right triangle)

2. **Determine the appropriate trigonometric function:**
- Since we know the opposite side ($\overline{ZY}$) and the hypotenuse ($\overline{XZ}$), we can use the sine function: $\sin \angle Z$ = opposite/hypotenuse = $\overline{ZY}/\overline{XZ}$ = $10/14$.

3. **Calculate the angle:**
- $\angle Z = \sin^{-1}(10/14) \approx 45.6°$

Answer: (C)

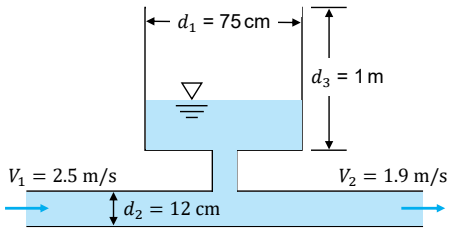
Qwen2.5-VL-32B (Split):
1. **Identify the given information:**
- $\overline{XZ}$ = 14 (hypotenuse), $\overline{ZY}$ = 10 (one leg)
- $\overline{ZY} \perp \overline{XY}$ (right triangle)

2. **Determine the appropriate trigonometric function:**
- We know the hypotenuse ($\overline{XZ}$), and $\angle Z$ is adjacent to the side ($\overline{ZY}$).
- The cosine function relates the hypotenuse to the adjacent side: $\cos \angle Z$ = adjacent/hypotenuse = $\overline{ZY}/\overline{XZ}$ = $10/14$.

3. **Calculate the angle:**
- $\angle Z = \cos^{-1}(10/14) \approx 44.4°$

Answer: (B)

2. Mechanical Engineering (Geometric/Physical Property Misinterpretation)



Question:
The pipe flow in the image fills a cylindrical surge tank as shown. At time t = 0, the water depth in the tank is 30 cm. Estimate the time required to fill the remainder of the tank.

Options:
(A) Δt = 50.614s (B) Δt = 45.614s
(C) Δt = 40.614s

Ground Truth: (B)

Qwen2.5-VL-32B (Integrated):
1. **Determine the volume of water that needs to be added:**
- The height to be filled is $d_1 - 0.3 = 0.75 - 0.3 = 0.45$ m.
- $V_0 = (d_3/2)^2\pi \times h = (1/2)^2\pi \times 0.45 \approx 0.3534$ m^3

2. **Calculate the net flow rate into the tank:**
- The cross-sectional area of the pipe at the inlet is $A_1 = (d_2/2)^2\pi = (0.12/2)^2\pi = 0.06^2\pi \approx 0.0113$ m^2
- The net flow rate into the tank is $Q_0 = (V_1 - V_2)A_1 = 0.6 \times 0.0113 \approx 0.0068$ m^3/s.

3. **Calculate the time required to fill the tank :**
- $\Delta t = V_0/Q_0 = 0.3534/0.0068 = 52.128$ s.

The closest option is (A) 50.614 s.

Qwen2.5-VL-32B (Split):
1. **Calculate the volume of the tank that needs to be filled:**
- The remaining height to be filled is $d_3 - 0.3 = 1 - 0.3 = 0.7$ m.
- The cross-sectional area of the tank is $A_T = (d_1/2)^2\pi \approx 0.4418$ m^2.
- The volume to be filled is $V_0 = A_T \times 0.7 \approx 0.3093$ m^3.

2. **Calculate the net flow rate into the tank:**
- The cross-sectional area of the inlet pipe is $A_1 = (d_2/2)^2\pi = (0.12/2)^2\pi = 0.06^2\pi \approx 0.0113$ m^2
- The net flow rate is $Q_0 = (V_1 - V_2)A_1 = 0.6 \times 0.0113 \approx 0.0068$ m^3/s.

3. ** Estimate the time required to fill the remainder of the tank:**
- $\Delta t = V_0/Q_0 = 0.3093/0.0068 = 45.614$ s.

Answer: (B)

Figure 4: Case studies of model predictions under integrated and split instructions in geometry and mechanical engineering. Incorrect reasoning steps are highlighted in red. For brevity, we present only the integrated instruction.

ments correctly but made mistakes in subsequent reasoning. Other types of errors include calculation errors, where the model correctly interpret visual elements, construct relevant equations, but fail to make a right computation.

Notably, these errors emerge under the integrated instruction setting, while the split instruction setting enables the model to correctly identify the relevant visual attributes and perform accurate reasoning. This suggests that physical integration of textual and visual information can make the model more vulnerable to errors whose types are well-known by the prior researches (Yue et al., 2024). Overall, these qualitative findings support the observations in Section 5.2 that physical integration is not universally beneficial for MLLMs.

6 Implications for MLLM Usage and Training

Our analysis of SPLINT provides valuable insights into the limitations of current MLLMs and suggests several promising directions for future MLLM usage and development.

Using Instructions Better Suited to MLLMs. As demonstrated in Table 2, using integrated instructions does not consistently improve MLLM

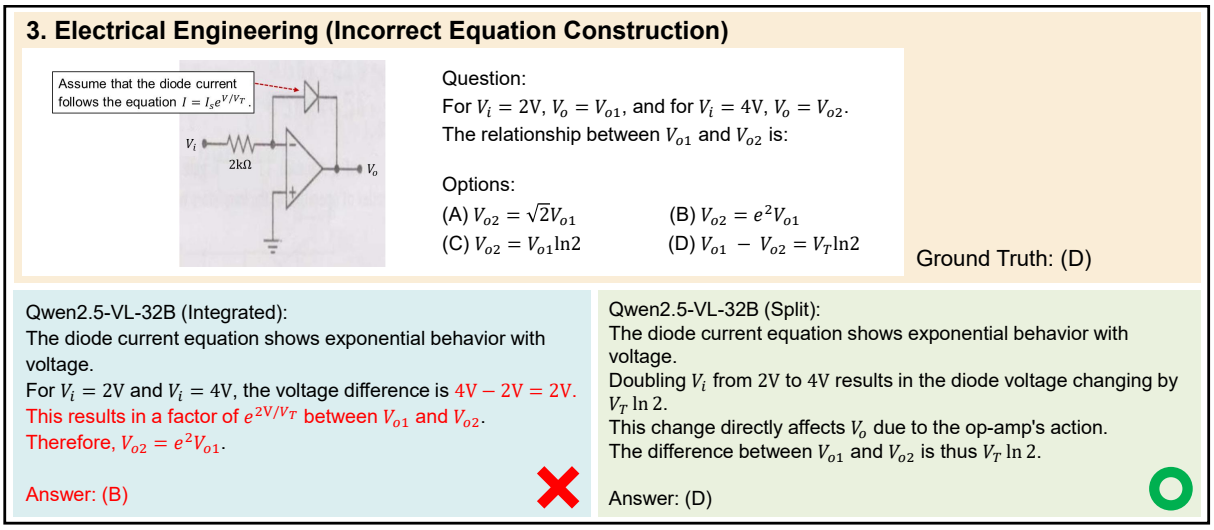


Figure 5: Case studies of model predictions under integrated and split instructions in electrical engineering. Incorrect reasoning steps are highlighted in red. For brevity, we present only the integrated instruction.

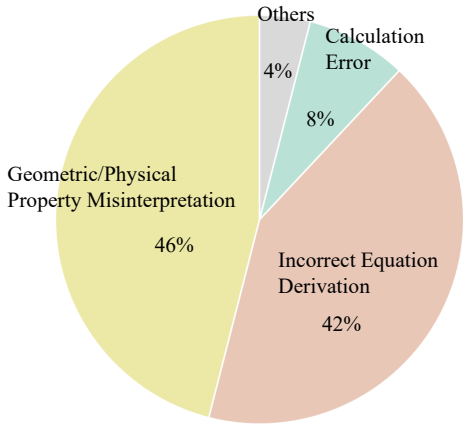


Figure 6: Error distribution over 100 annotated MLLM errors on integrated instructions.

reasoning and even decrease the performance in majority of cases. In addition, notable portion of these cases exhibits significant performance degradation on integrated instruction settings. These results indicate a potential risk that physical integration originally intended to facilitate human reasoning may not work for MLLMs. This trend highlights that physical integration should be applied with caution when designing instructions for MLLMs, and it may be preferable to avoid using integrated instructions for better MLLM reasoning.

Constructing CoT Training Data for Physically Integrated Instructions. As observed in Figure 4, integrated instructions can induce diverse failure modes in the CoT reasoning processes of MLLMs, and these errors often propagate throughout subsequent reasoning. Recent studies have shown that fine-tuning small MLLMs on CoT data is a promising paradigm for improving visual reasoning capabilities (Zhang et al., 2024b). However, these works have paid limited attention to how the arrangement of textual and visual information within an image affects CoT reasoning. Therefore, a promising direction for future research is to construct CoT training data involving physically integrated instructions, enabling MLLMs to better reason under such settings.

7 Conclusion

In this work, we introduce SPLINT, a novel benchmark designed to investigate whether physical integration of textual and visual information improves reasoning in MLLMs. Motivated by cognitive load theory, SPLINT consists of 925 paired visual question-answer instances under both integrated and split instruction settings. Using this benchmark, we evaluate the reasoning capabilities of 12 popular MLLMs across the two settings. Our experiments reveal that physical integration of textual and visual information does not consistently enhance reasoning performance and can even degrade performance in certain cases. Further qualitative analysis suggests that physical integration of textual and visual information can make the model more vulnerable to errors whose types are well-known by the prior researches. These findings suggest that physical integration is not universally beneficial for MLLMs and should therefore be applied with careful consideration. We hope that SPLINT will serve as a valuable benchmark for future research on improving reasoning in MLLMs.

Limitations

Our work has two main limitations that should be addressed in future research.

(1) We primarily focus on three domains-geometry, mechanical engineering, and electrical engineering-when constructing SPLINT. These domains are selected because prior cognitive science research has shown that physical integration of textual and visual information improves human reasoning in these domains (Chandler and Sweller, 1991). However, they are not the only domains in which physical integration has been shown to be beneficial for human reasoning, and physical integration may systematically affect MLLM reasoning in those domains as well. Extending SPLINT to additional domains would therefore be an interesting direction for future work.

(2) Currently, we mainly evaluate MLLMs on SPLINT using direct prompting and CoT prompting. However, MLLMs can utilize many other advanced prompting strategies for improved reasoning, including in-context learning (Brown et al., 2020) and decomposition-based prompting methods (Wang et al., 2023). Evaluating MLLMs under these advanced prompting settings is left for future work.

References

David Acuna, Ximing Lu, Jaehun Jung, Hyunwoo Kim, Amlan Kar, Sanja Fidler, and Yejin Choi. 2025. Socratic-mcts: Test-time visual reasoning by asking the right questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 24147–24160. Association for Computational Linguistics.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Paul Chandler and John Sweller. 1991. Cognitive load theory and the format of instruction. *Cognition and instruction*, 8(4):293–332.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 8199–8221. Association for Computational Linguistics.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. 2024. Agent AI: surveying the horizons of multimodal interaction. *CoRR*, abs/2401.03568.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. BLINK: multimodal large language models can see but not perceive. In *Computer Vision - ECCV 2024 - 18th European Conference*, Lecture Notes in Computer Science, pages 148–166. Springer.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. 2025. Interleaved-modal chain-of-thought. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, pages 19520–19529. Computer Vision Foundation / IEEE.

Nitzan Bitton Guetta, Aviv Slobodkin, Aviya Maimon, Eliya Habba, Royi Rassin, Yonatan Bitton, Idan Szpektor, Amir Globerson, and Yuval Elovici. 2024. Visual riddles: a commonsense and world knowledge challenge for large vision and language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

Meng-Hao Guo, Xuanyu Chu, Qianrui Yang, Zhe-Han Mo, Yiqing Shen, Pei-lin Li, Xinjie Lin, Jinnian Zhang, Xin-Sheng Chen, Yi Zhang, Kiyohiro Nakayama, Zhengyang Geng, Houwen Peng, Han Hu, and Shi-Min Hu. 2025. Rbench-v: A primary assessment for visual reasoning models with multi-modal outputs. *CoRR*, abs/2505.16770.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. 2024. Olympiadbench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 3828–3850. Association for Computational Linguistics.

Yushi Hu, Weijia Shi, Xingyu Fu, Dan Roth, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Ranjay Krishna. 2024. Visual sketchpad: Sketching as a

visual chain of thought for multimodal language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

Amirmohammad Izadi, Mohammadali Banayeeanzade, Fatemeh Askari, Ali Rahimiakbar, Mohammad Mahdi Vahedi, Hosein Hasani, and Mahdieh Soleymani Baghshah. 2025. Visual structures help visual reasoning: Addressing the binding problem in LVLMs. In *Advances in Neural Information Processing Systems 39: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025*.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024a. Mantis: Interleaved multi-image instruction tuning. *Trans. Mach. Learn. Res.*, 2024.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, Bo Zhang, Chaoyou Fu, Peng Gao, and Hongsheng Li. 2025. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. In *Forty-second International Conference on Machine Learning, ICML 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Xinyi Jiang, Guoming Wang, Junhao Guo, Juncheng Li, Wenqiao Zhang, Rongxing Lu, and Siliang Tang. 2024b. DIEM: decomposition-integration enhancing multimodal insights. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 27294–27303. IEEE.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *AAdvances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc.

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

Danrui Li, Jiahao Zhang, Bernhard Egger, Moitreya Chatterjee, Suhas Lohit, Tim K Marks, and Anoop Cherian. 2026. Assemblybench: Physics-aware assembly of complex industrial objects. *arXiv preprint arXiv:2605.12845*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024b. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024*, pages 14369–14387. Association for Computational Linguistics.

Ming Li, Jike Zhong, Tianle Chen, Yuxiang Lai, and Konstantinos Psounis. 2025a. Eee-bench: A comprehensive multimodal electrical and electronics engineering benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 13337–13349. Computer Vision Foundation / IEEE.

Xiujun Li, Yujie Lu, William Yang Wang, and Yejin Choi. 2024c. Pixelated instructions: Can multimodal large language models follow printed instructions in images? In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*.

Yunxin Li, Zhenyu Liu, Zitao Li, Xuanyu Zhang, Zhenran Xu, Xinyu Chen, Haoyuan Shi, Shenyuan Jiang, Xintong Wang, Jifang Wang, Shouzheng Huang, Xinping Zhao, Borui Jiang, Lanqing Hong, Longyue Wang, Zhuotao Tian, Baoxing Huai, Wenhan Luo, Weihua Luo, and 3 others. 2025b. Perception, reason, think, and plan: A survey on large multimodal reasoning models. *CoRR*, abs/2505.04921.

Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, and 1 others. 2024d. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. In *AI for Accelerated Materials Design-Vienna 2024*.

Zhiyuan Li, Dongnan Liu, Chaoyi Zhang, Heng Wang, Tengfei Xue, and Weidong Cai. 2024e. Enhancing advanced visual reasoning ability of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pages 1915–1929. Association for Computational Linguistics.

Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, Yuheng Cheng, Suyuchen Wang, Xiaoqiang Wang, Yuyu Luo, Haibo Jin, Peiyan Zhang, Ollie Liu, Jiaqi Chen, Huan Zhang, and 28 others. 2025. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *CoRR*, abs/2504.01990.

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

Xu Liu, Yongheng Zhang, Qiguang Chen, Yao Li, Sheng Wang, and Libo Qin. 2026. Let's think with images efficiently! an interleaved-modal chain-of-thought reasoning framework with dynamic and precise visual thoughts. In *Fortieth AAAI Conference on Artificial Intelligence, AAAI 2026*, pages 32213–32221. AAAI Press.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV*

2024 - 18th European Conference, Lecture Notes in Computer Science, pages 216–233. Springer.

Yuxing Long, Jiyao Zhang, Mingjie Pan, Tianshu Wu, Taewhan Kim, and Hao Dong. 2025. Checkmanual: A new challenge and benchmark for manual-based appliance manipulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025*, pages 22595–22604. Computer Vision Foundation / IEEE.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021*, pages 6774–6786. Association for Computational Linguistics.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. 2024. Compositional chain-of-thought prompting for large multimodal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 14420–14431. IEEE.

Atsuyuki Miyai, Shota Onohara, Jeonghun Baek, and Kiyoharu Aizawa. 2025. Jmmmu-pro: Image-based japanese multi-discipline multimodal understanding benchmark via vibe benchmark construction. *CoRR*, abs/2512.14620.

OpenAI. 2024. Hello gpt-4o.

OpenAI. 2025. Introducing gpt-5.

John Sweller, Paul Chandler, Paul Tierney, and Martin Cooper. 1990. Cognitive load as a factor in the structuring of technical material. *Journal of experimental psychology: general*, 119(2):176.

John Sweller, Jeroen JG Van Merrienboer, and Fred GWC Paas. 1998. Cognitive architecture and instructional design. *Educational psychology review*, 10(3):251–296.

Gemma Team. 2025a. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Qwen Team. 2025b. Qwen3-vl technical report. *CoRR*, abs/2511.21631.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

Yael Vinker, Tamar Rott Shaham, Kristine Zheng, Alex Zhao, Judith E. Fan, and Antonio Torralba. 2025. Sketchagent: Language-driven sequential sketch generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 23355–23368. Computer Vision Foundation / IEEE.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pages 2609–2634. Association for Computational Linguistics.

Yikun Wang, Siyin Wang, Qinyuan Cheng, Zhaoye Fei, Liang Ding, Qipeng Guo, Dacheng Tao, and Xipeng Qiu. 2025. Visuothink: Empowering LVLM reasoning with multimodal tree search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, ACL 2025*, pages 21707–21719. Association for Computational Linguistics.

Jinyang Wu, Mingkuan Feng, Guocheng Zhai, Shuai Zhang, Zheng Lian, Fangrui Lv, Pengpeng Shao, Ruihan Jin, Zhengqi Wen, and Jianhua Tao. 2026. Astar: Boosting multimodal reasoning with automated structured thinking. In *Fortieth AAAI Conference on Artificial Intelligence, AAAI 2026*, pages 33926–33934. AAAI Press.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

Kun Xiang, Heng Li, Terry Jingchen Zhang, Yinya Huang, Zirong Liu, Peixin Qu, Jixi He, Jiaqi Chen, Yu-Jie Yuan, Jianhua Han, Hang Xu, Hanhui Li, Mrinmaya Sachan, and Xiaodan Liang. 2025. Seephys: Does seeing help thinking? - benchmarking vision-based physics reasoning. *CoRR*, abs/2505.19099.

Ziyang Xiao, Dongxiang Zhang, Xiongwei Han, Xiaojin Fu, Wing Yin Yu, Tao Zhong, Sai Wu, Yuan Wang, Jianwei Yin, and Gang Chen. 2024. Enhancing LLM reasoning via vision-augmented prompting. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024*.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang,

Yu Qiao, and Ping Luo. 2025. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(3):1877–1893.

Huanjin Yao, Jiaxing Huang, Yawen Qiu, Michael K. Chen, Wenzheng Liu, Wei Zhang, Wenjie Zeng, Xikun Zhang, Jingyi Zhang, Yuxin Song, Wenhao Wu, and Dacheng Tao. 2025. Mmreason: An open-ended multi-modal multi-step reasoning benchmark for mllms toward AGI. *CoRR*, abs/2506.23563.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Zhenfei Yin, Jiong Wang, Jianjian Cao, Zhelun Shi, Dingning Liu, Mukai Li, Xiaoshui Huang, Zhiyong Wang, Lu Sheng, Lei Bai, Jing Shao, and Wanli Ouyang. 2023. LAMM: language-assisted multi-modal instruction-tuning dataset, framework, and benchmark. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Haoxuan You, Rui Sun, Zhecan Wang, Long Chen, Gengyu Wang, Hammad A. Ayyubi, Kai-Wei Chang, and Shih-Fu Chang. 2023. Idealgpt: Iteratively decomposing vision and language reasoning via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Findings of ACL, pages 11289–11303. Association for Computational Linguistics.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024*, Proceedings of Machine Learning Research, pages 57730–57754. PMLR / OpenReview.net.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pages 9556–9567. IEEE.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics, ACL 2025*, pages 15134–15186. Association for Computational Linguistics.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 12401–12430. Association for Computational Linguistics.

Liang Zhang, Anwen Hu, Jing Zhang, Shuo Hu, and Qin Jin. 2023. MPMQA: multimodal question answering on product manuals. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*, pages 13958–13966. AAAI Press.

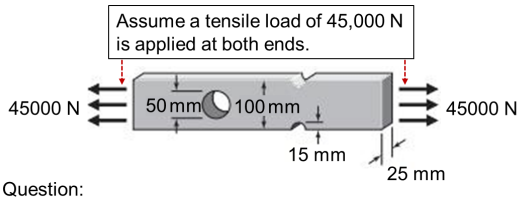
Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025*, pages 17733–17750. Association for Computational Linguistics.

Kaiwen Zhou, Kwonjoon Lee, Teruhisa Misu, and Xin Wang. 2024. Vicor: Bridging visual understanding and commonsense reasoning with large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 10783–10795. Association for Computational Linguistics.

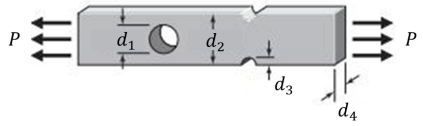
1. Question w/ Physical Integration



Question:
Find the maximum value of stress at the hole and semicircular notch shown in Figure.

Options:
(A) 71.36 Mpa, 40.65 Mpa **(B) 81.36 Mpa, 50.65 Mpa**
(C) 71.36 Mpa, 30.65 Mpa (D) 61.36 Mpa, 50.65 Mpa

2. Question w/o Physical Integration



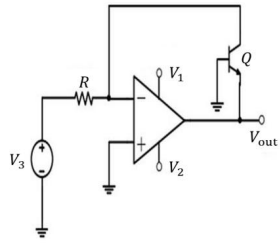
Question:
Find the maximum value of stress at the hole and semicircular notch shown in Figure. Assume a tensile load of P = 45,000 N is applied at both ends. Also assume d_1 = 50 mm, d_2 = 100 mm, d_3 = 15 mm, and d_4 = 25 mm.

Options:
(A) 71.36 Mpa, 40.65 Mpa **(B) 81.36 Mpa, 50.65 Mpa**
(C) 71.36 Mpa, 30.65 Mpa (D) 61.36 Mpa, 50.65 Mpa

Figure 7: Comparison of (a) integrated and (b) split instructions for a mechanical engineering problem.Green text indicates the correct answer.

A More Examples of Instructions with/without Physical Integration

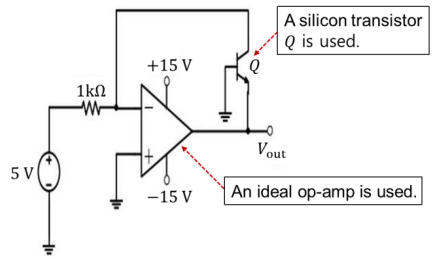
1. Question w/ Physical Integration



Question:
In the circuit shown in the image, assume $V_1 = +15$ V, $V_2 = -15$ V, $V_3 = 5$ V, and $R = 1\text{k}\Omega$. What is the output voltage V_{out} if a silicon transistor Q and an ideal op-amp are used?

Options:
(A) -15 V **(B) -0.7 V** (C) $+0.7$ V (D) $+15$ V

2. Question w/o Physical Integration



Question:
In the circuit shown in the image, what is the output voltage V_{out}?

Options:
(A) -15 V **(B) -0.7 V** (C) $+0.7$ V (D) $+15$ V

Figure 8: Comparison of (a) integrated and (b) split instructions for a electrical engineering problem.Green text indicates the correct answer.

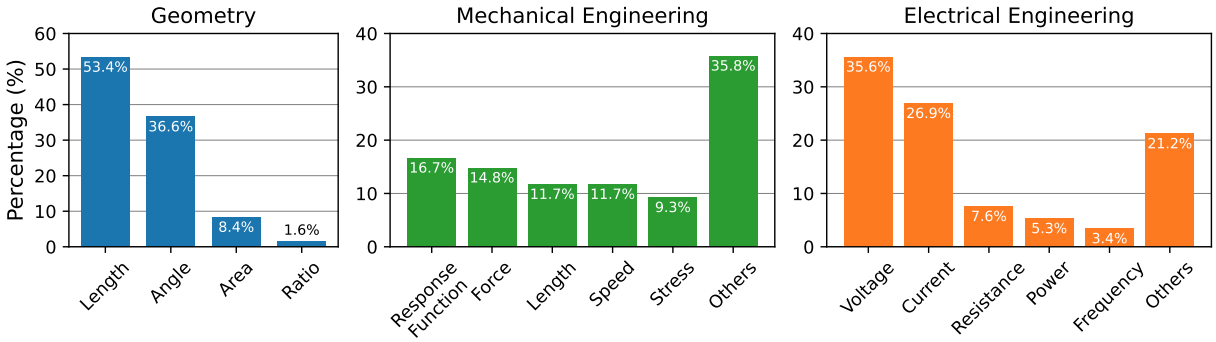


Figure 9: Distributions of answer types across domains. For each domain, at most five of the most frequent types are shown.

System Prompt
You are a helpful assistant.
User Prompt
Direct → Output only one of the choices. Do not include any explanation. **CoT** → Answer the multiple choice question given in the image. The last line of your response should be of the following format: 'Answer: $LETTER' (without quotes) where LETTER is one of the options. Think step by step before answering.

Table 3: Prompts we used in our experiments.