



UvA-DARE (Digital Academic Repository)

Same Content, Different Answers: Cross-Modal Inconsistency in MLLMs

van Sprang, A.; Samson, L.; Lucic, A.; Acar, E.; Ghebreab, S.; Asano, Y.M.

DOI

[10.48550/arXiv.2512.08923](https://doi.org/10.48550/arXiv.2512.08923)

Publication date

2025

Document Version

Submitted manuscript

License

CC BY

[Link to publication](#)

Citation for published version (APA):

van Sprang, A., Samson, L., Lucic, A., Acar, E., Ghebreab, S., & Asano, Y. M. (2025). *Same Content, Different Answers: Cross-Modal Inconsistency in MLLMs*. ArXiv. <https://doi.org/10.48550/arXiv.2512.08923>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

Same Content, Different Answers: Cross-Modal Inconsistency in MLLMs

Angela van Sprang^{1*}, Laurens Samson^{1,2*},

Ana Lucic¹, Erman Acar¹, Sennay Ghebrea¹, Yuki M. Asano³

¹University of Amsterdam

²City of Amsterdam

³University of Technology Nuremberg

Abstract

*Multimodal large language models (MLLMs) are trained to represent vision and language in a shared space. But does this joint representation enable consistent reasoning across modalities? We introduce **REST** and **REST+** (Render-Equivalence Stress Tests), two benchmarks for systematically evaluating cross-modal consistency. Each sample presents **semantically identical information** in three forms (image, text, and mixed), allowing us to measure whether models produce consistent outputs regardless of modality. Evaluating 15 state-of-the-art MLLMs, we find that none reason consistently across modalities, with substantial variation in the degree of inconsistency. Neither rendering text as images nor images as text resolves this problem, even when controlling for OCR errors. We further show that visual characteristics (color, resolution, but not font) and the number of vision tokens affect performance even when text is correctly recognized. Finally, our consistency score correlates with the cross-modal cosine similarity in embedding space, suggesting a mechanistic explanation: inconsistent reasoning arises when text and image representations occupy distinct regions of the joint space*

1. Introduction

Multimodal large language models (MLLMs) [3, 4, 11, 21, 22, 43], have achieved remarkable progress across diverse multimodal tasks, demonstrating strong capabilities in visual question answering [23, 44, 45], document understanding [25, 26], and complex reasoning [4, 21, 31]. The success of MLLMs suggests they have learned to integrate visual and textual information for multimodal understanding seamlessly, even though these modalities might not explicitly be integrated. In particular, recent work points out the existence of a *modality gap* in different MLLMs, where text and image embeddings occupy different regions in the joint embedding space [20, 30, 34]. The smaller this gap, the better the downstream task performance [13, 34].

Simultaneously, recent work by the authors of DeepSeek-OCR [40] explores the idea of rendering text as images to reduce token costs. They demonstrate that compressing 10 text tokens into a single visual token maintains 97% optical character recognition (OCR) accuracy, offering substantial computational savings for text-heavy inputs. However, this raises a fundamental question given the modality gap: *when a model successfully reads text from an image, does it reason about that information as effectively as when receiving native text?* If not, this could pose critical problems: MLLMs might perform better on the text modality, as their foundation is an LLM. They might produce incorrect answers based solely on the prompt modality, rather than on the information it contains. We refer to this as **cross-modal inconsistency**: obtaining inconsistent results on prompts with semantically identical (“same content”) information in different modalities.

Existing work has documented cross-modal inconsistencies in MLLMs [5, 32, 33, 35, 45], but prior benchmarks either evaluate only a single model [46] or do not control for actual readability (OCR, [8]), potentially conflating text recognition failures with reasoning inconsistencies.

To address this gap, we introduce **REST** (Render-Equivalence Stress Tests) and **REST+** to systematically evaluate cross-modal reasoning consistency under controlled OCR difficulty. Both benchmarks present the same semantic content in three formats: text-only, image-only, and mixed (question in text, context in image). **REST** measures performance on SOEBENCH, our newly generated system-of-equations task with easy-to-recognise symbols, and on three established benchmarks where text is rendered as an image (MMLU [16], ARC [9], GSM-Symbolic [27]). SOEBENCH ensures minimal OCR burden and guarantees zero prior exposure, reducing potential memorisation issues by MLLMs. All analyses are performed conditional on the model producing correct OCR for the given question. **REST+** extends **REST** by permuting each image in 10 different ways, varying the font, resolution, and text color. This allows us to evaluate the consistency of reasoning skills across formats, including the amount of visual tokens. Our research questions are:

¹Shared first author. {a.v.vansprang, l.samson}@uva.nl

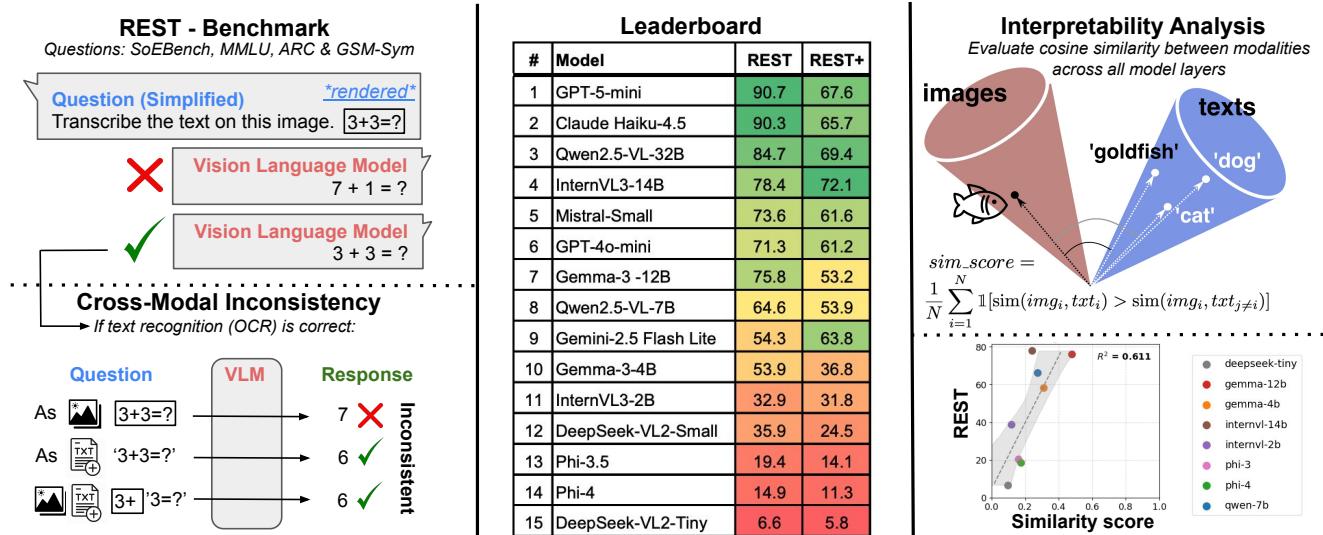


Figure 1. **Summary of our work.** *Left:* Our **REST** benchmark measures whether MLLMs can consistently reason over identical information across modalities. We first verify text recognition (OCR) capability, then evaluate the same question in three modalities (text, image, mixed). Cross-modal inconsistency occurs when models produce different answers depending on the input format. *Center:* RER consistency score measures the degree to which a model outputs the same answer in all modalities. We evaluate 15 MLLMs (on OCR correct, sorted on **REST**) and find that the degree of cross-modal inconsistency varies substantially across models even when controlling for OCR. *Right:* Matching samples (i.e., different modalities containing the same information) show higher cosine similarity than non-matching ones, and the extent of this difference correlates with the consistency score on our benchmark.

RQ1: Do current frontier MLLMs exhibit cross-modal inconsistency? If so, which modalities perform best?

RQ2: Is cross-modal inconsistency due to OCR performance rather than true modality differences?

RQ3: Do visual characteristics such as resolution (**RQ3a**) or font and text colour (**RQ3b**) influence cross-modal inconsistency, and if so, how?

RQ4: Given the internal representations of the same concept in different modalities, is there a correlation between the cross-modal similarity of internal representations and the degree of cross-modal inconsistency?

We have three main contributions. First, we introduce **REST** and **REST+** as benchmarks to measure cross-modal inconsistency. Unlike previous benchmarks, we include a new set of tasks (SOEBENCH) that is guaranteed to not be seen during pre-training and control for OCR complexity. Next, we evaluate 15 frontier MLLMs and find substantial inconsistencies across modalities (at least $\sim 10\%$ inconsistency), even when controlling for OCR. This leaves a notable gap of solvable questions that current models fail to capture. Finally, we analyse the internal representations of matching samples (i.e., samples with the same information in different modalities) and find that they show higher cosine similarity than non-matching pairs, and that this similarity magnitude correlates with our consistency score. We will release all code and data upon publication.

2. Our Benchmarks: REST and REST+

We introduce our **REST** and **REST+** benchmarks to enable the systematic evaluation of cross-modal inconsistency in MLLMs. We investigate whether models consistently reason over identical information presented across different modalities, while controlling for potential confounding effects from OCR performance.

2.1. REST

Our benchmark consists of four distinct evaluation tasks:

- **OCR:** Extract rendered text from the image.
- **Text:** Answer textual questions.
- **Image:** Answer questions rendered as images.
- **Mixed:** Answer questions with the context as image and the question as text (or, in case of multiple-choice questions, the question is the image and the options are text).

Our benchmark consists of our SOEBENCH task and three established benchmarks: MMLU [15], ARC [10], and GSM-Symbolic [27]. To isolate cross-modal consistency from OCR failures, we minimise OCR complexity (i.e. we prevent models from failing to understand the images) by excluding questions with more than 800 characters (threshold decided by initial experiments) and questions with LaTeX code (for which OCR is ambiguous). To keep the text easy to read, we render images with white backgrounds and high-resolution (DPI 200) black DejaVu Sans text.

2.2. REST+

We introduce **REST+** to study the effect of visual characteristics (such as resolution, font and colour) on cross-modal inconsistency. **REST+** increases complexity by creating 10 visual permutations of each image while preserving semantic content. For computational feasibility, we only include the text and image tasks. We use MMLU [15] with the same question filtering as in **REST**. We create 10 permutations for each question: 9 combinations from 3 font families (DejaVu Sans, Courier New, Cursive), 3 DPI values (50, 100, 200), and one coloured variant (red, green, blue, cyan, magenta, or yellow) at 200 DPI using DejaVu Sans. We manually verified that the text is legible at 50 DPI for each font. We randomly sample 10% per subject class for computational feasibility, which results in 1085 final questions. Sample images are provided in the Appendix.

2.3. SOEBENCH

SOEBENCH is task suite based on solving systems of linear equations presented in text, image, or mixed modalities included in **REST** and **REST+**. Each question contains symbolic variables and requires basic algebraic reasoning, with all instances generated to have a single integer solution. Its restricted symbol set (digits 0–9 and letters A–E) reduces OCR complexity, ensuring that performance differences reflect reasoning rather than recognition errors. Since it is newly generated, we can guarantee existing MLLMs have had no prior exposure to it.

3. Benchmarking Methodology

For each **REST** question, the MLLM receives four input prompts. First, we verify OCR performance by letting the model transcribe the image, ensuring that the model can read the question. For the remaining tasks (text, mixed, image), we employ Chain-of-Thought (CoT) prompting. All prompts are shown in the Appendix. We limit the output to 1024 tokens for all tasks, except for SOEBENCH, for which we allow 2048 output tokens because 1024 is not always sufficient. We set the temperature to 0 for reproducibility. We parse answers using regular expressions and treat format-invalid responses as incorrect.

We introduce three metrics to evaluate our benchmark on cross-modal inconsistency. We measure answer consistency across modalities with the *Render-Equivalence Rate* (RER), which is similar to the consistency score from Zhang et al. [46]:

$$\text{RER} = \frac{|\{x \mid f(x_t) = f(x_i, z_i) = f(x_m, z_m)\}|}{N}, \quad (1)$$

where x denotes text prompts and z denotes images for input type (text, image or mixed), f is the model, and N is the dataset size. This metric represents the fraction of questions

for which the model gives the same answer for all modalities, where 1 denotes a perfectly consistent model.

We also evaluate inconsistency in questions solvable only through specific modalities, i.e., those answerable in at least one, but not all, modalities. We use the *Cross-Modality Failure Rate* (CFR):

$$\text{CFR} = \frac{|\{q \mid 1 \leq \sum_{m \in M} C(q, m) < |M|\}|}{N_c}, \quad (2)$$

where N_c is the number of questions that are answered correctly in at least one modality, $M = \{\text{text, image, mixed}\}$, and $C(q, m)$ indicates correctness (1 correct, 0 incorrect) for question q and modality m . We exclude questions that failed for all modalities since these represent inability rather than inconsistency. A score of 0 indicates a perfectly consistent model.

We assess overall capability using *Max Modal Coverage* (MMC):

$$\text{MMC} = \frac{|\{q \mid 1 \leq \sum_{m \in M} C(q, m)\}|}{N}, \quad (3)$$

which represents the fraction of questions solved through at least one modality. A score of 1 indicates (potentially) perfect performance on all questions, provided that each question is presented in the suitable modality.

To evaluate OCR performance, we use *Character Error Rate* (CER), which measures character insertions, deletions, and substitutions and normalises by reference length. For the evaluation of cross-modality consistency, we include only the questions where text is perfectly recognised. We only evaluate letters and digits to focus on semantic recognition rather than punctuation.

4. Benchmarking Results: REST

4.1. RQ1: Cross-modal inconsistency in MLLMs

Table 1 presents RER and CFR scores for 15 state-of-the-art MLLMs, with separate scores for all data and only those questions with correct OCR (*OCR✓*). The scores are averaged over all four **REST** benchmarks. The RER consistency scores vary substantially across MLLMs, ranging from 6.6% to 90.7%. Notably, RER scores are equal or slightly higher for all models when only evaluating questions with correct OCR. This indicates that OCR errors do affect consistency, and our efforts to control OCR complexity are justified. Moreover, the model rankings are not affected by excluding incorrect OCR questions. Overall, closed-source models (GPT-5-mini [29] and Claude Haiku 4.5 [6]) achieve the highest consistency scores (90.7% and 90.3%, respectively, when OCR is correct). Gemini 2.5 Flash Lite [37] and GPT-4o-mini [37] perform worse than other closed-source models. Among open-source models, Qwen-2.5 32B [7] leads with an RER of 84.7%.

Table 1. **REST scores for 15 MLLMs show varying degrees of cross-modal inconsistency.** RER and CFR scores are given for all questions and for those where models correctly recognised the rendered text (OCR \checkmark). All models achieve equal or higher RER scores on the OCR \checkmark subset, though differences are minor.

Model	REST (OCR \checkmark)		REST (All Questions)		
	RER \uparrow	CFR \downarrow	RER \uparrow	CFR \downarrow	OCR \checkmark \uparrow
	Deepseek-Tiny [41]	6.6	98.0	6.5	98.1
Phi-4 [2]	14.9	82.3	14.5	82.8	94.0
Phi-3.5 [1]	19.4	79.3	19.3	79.5	90.3
InternVL3 (2B) [47]	32.9	63.7	32.6	63.9	90.7
Deepseek-Small [41]	35.9	60.9	35.9	60.9	97.9
Gemma-3 (4B) [38]	53.9	42.3	52.3	44.0	83.5
Gemini-2.5 Fl. Lite [37]	54.3	40.3	54.1	40.4	98.3
Qwen-2.5 (7B) [7]	64.6	31.7	64.3	31.9	98.8
GPT-4o-mini [18]	71.3	26.0	71.2	26.1	98.8
Mistral-Small [28]	73.6	23.9	73.4	24.1	98.4
Gemma-3 (12B) [38]	75.8	21.3	75.5	21.6	96.5
InternVL3 (14B) [47]	78.4	19.6	78.1	19.9	95.3
Qwen-2.5 (32B) [7]	84.7	13.6	84.5	13.7	97.5
Haiku-4.5 (Claude) [6]	90.3	8.9	90.1	9.1	98.2
GPT-5-mini [29]	90.7	8.7	90.7	8.8	99.0

The CFR results reveal concerning patterns: even GPT-5-mini fails to consistently solve 8.7% of questions across all modalities despite solving them in at least one modality. This inconsistency even reaches 82.3% for Phi-4 [2], which means that obtaining a correct answer depends on how the input is formatted (text, mixed, or image).

Max modal coverage Given the presence of cross-modal inconsistency in MLLMs, we examine which modalities obtain more correct answers. Figure 2 shows models with low, medium, and high consistency scores for GSM8k-Symbolic [27]. We report scores on this dataset because it contains open-ended questions that cannot be solved by guessing. The results show that models with low consistency could increase model performance drastically when solving questions in the optimal modality. In particular, MMC (i.e., fraction of questions that are solvable through at least one modality) for Phi-4 is 85.9%, but only 36.7% of questions can be solved in all modalities. So, 49.2% of questions might obtain incorrect answers if they are not in the optimal modality. Similarly, DeepSeek-VL-Small [41] can only solve 10.2% of the questions through text alone. Qwen2.5-VL-32B shows the best consistency, where almost all questions can be solved in all modalities.

Preference for text modality Using the different tasks in REST, we find that almost all models perform better in the text modality (see Table 2 for the accuracies on MMLU, AI2-ARC and GSM8k-Symbolic). With a statistical t-test, and combining the answers from all benchmarks, we find

Table 2. **Preference for Text.** This table presents accuracies for the text, mixed, and image tasks on MMLU, AI2-ARC, and GSM8k-Symbolic (OCR-correct subset). Bold values indicate the best-performing modality for each model-dataset combination.

Model	MMLU [16] \uparrow			AI2-ARC [9] \uparrow			GSM-Sym [27] \uparrow		
	Text	Mix	Img	Text	Mix	Img	Text	Mix	Img
Deepseek-T	29.5	26.6	25.3	34.2	28.7	27.5	15.5	0.8	0.7
Deepseek-S	54.9	54.9	51.2	69.1	68.7	70.0	67.3	54.7	60.9
GPT-4o-mini	84.4	77.2	77.0	93.6	89.3	89.9	91.4	86.7	89.0
GPT-5-mini	89.3	86.5	87.8	94.6	92.7	93.3	95.1	93.7	94.1
Gemini-2.5 FL	81.6	79.7	78.4	90.4	89.0	88.8	79.4	60.1	50.5
Gemma-3-4B	68.0	64.7	57.2	79.2	78.2	71.4	82.2	70.5	67.4
Gemma-3-12B	77.6	76.7	72.2	90.5	90.2	88.6	91.3	87.7	88.1
Haiku-4.5	90.1	87.3	85.1	93.3	93.3	92.7	95.6	94.9	94.1
InternVL3-2B	61.6	59.0	56.6	76.0	74.0	74.0	52.1	40.3	53.1
InternVL3-14B	82.9	81.4	83.0	92.9	92.9	94.2	91.2	91.9	93.0
Mistral-Small	84.1	78.9	82.5	92.7	90.0	92.6	92.6	91.7	93.3
Phi-3.5	61.1	41.5	32.5	73.3	54.3	37.9	66.9	42.9	48.3
Phi-4	47.0	44.2	31.4	59.9	48.5	35.0	76.5	55.8	54.5
Qwen-2.5-7B	72.6	71.8	72.9	86.1	87.0	88.3	84.4	82.0	82.1
Qwen-2.5-32B	83.3	84.1	83.5	92.9	92.9	92.9	93.6	93.1	93.0

that models prefer text over image ($t = 17.7$, $p < 0.05$) and that image is more difficult than mixed ($t = -7.2$, $p < 0.05$). Several model families (Phi, Gemma, ChatGPT, and Claude) consistently achieve their best performance in text rather than other modalities (e.g., GPT-4o-mini obtains more than 7% higher accuracy in text than in mixed or image on MMLU).

This raises the question of whether we observe data contamination, where the questions of these text benchmarks were included in the models’ training data, or whether the text modality is inherently more effective in MLLMs.

4.2. RQ2: Controlling for OCR

Therefore, we evaluate cross-modal inconsistency on data that all models have not seen before, while keeping OCR simple, by using our SOEBENCH. Table 3 shows the consistency scores, task accuracies and OCR accuracies on this benchmark. Firstly, all models can solve OCR (near) perfectly, with the exception of DeepSeek-Tiny, which reproduces the example answer from the input prompt. Moreover, the Gemma and GPT models still outperform on the text modality. Surprisingly, Phi-4 now performs better on the image modality, while the other benchmarks showed higher performance on the text modality by a large margin.

We argue that OCR and data contamination are not the main reasons for cross-modal inconsistency, as models inherently perform better through the text modality.

OCR first, then solve Chen et al. [8] demonstrate improved results when instructing MLLMs to first extract information from an image before solving the task. Table 4 reports the accuracy difference between direct image pro-

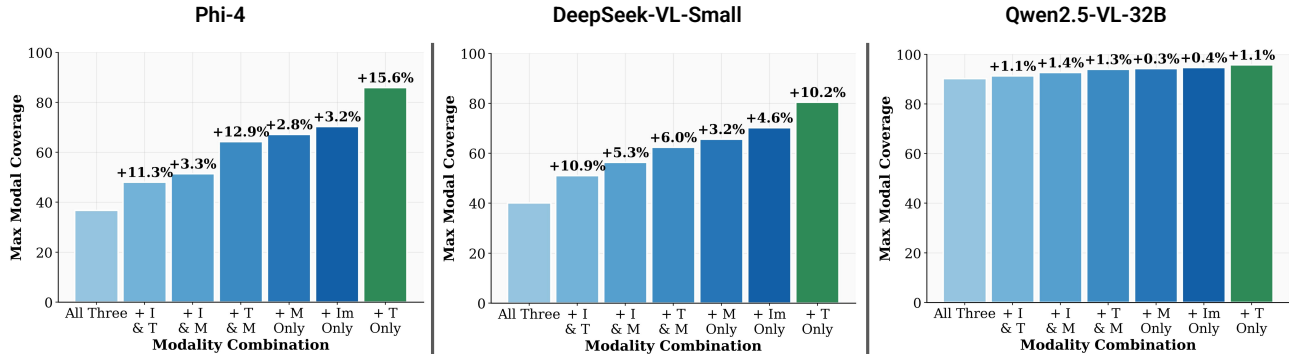


Figure 2. **Cross-modal inconsistency leaves model potential untapped.** This figure shows the cumulative distribution of correctly solved questions across sets of modalities (OCR-correct subset). From left to right, the bars represent: the percentage of questions that can be solved in all three modalities, followed by including questions that can only be solved in fewer modalities, ending with the Max Modal Coverage (green), which shows the percentage of questions that can be solved in at least one modality. Results are shown for GSM8k-Symbolic, which contains open-ended questions that cannot be solved by guessing.

Table 3. **Models inherently perform better through the text modality.** These results on our SOEBENCH include consistency scores, accuracies for the different tasks and OCR accuracy (OCR-correct subset). Even when the data is novel and models achieve (nearly) perfect OCR scores, the text modality is still preferred.

Model	RER \uparrow	Text \uparrow	Mixed \uparrow	Image \uparrow	OCR \uparrow
Phi-3.5	0.0	1.3	1.3	0.7	100.0
Deepseek-Tiny	0.0	0.7	0.0	0.0	0.0
Deepseek-Small	0.7	8.0	7.3	4.7	100.0
Phi-4	1.3	13.3	13.3	17.3	100.0
InternVL3 (2B)	4.7	14.0	14.7	33.3	100.0
Gemini-2.5 Flash Lite	13.3	42.7	34.7	49.3	100.0
Qwen-2.5 (7B)	24.7	48.0	46.0	48.7	100.0
GPT-4o-mini	39.3	65.3	61.3	62.0	100.0
Mistral-Small	41.3	60.7	62.0	62.0	100.0
Gemma-3 (4B)	41.9	75.2	58.9	57.4	86.0
InternVL3 (14B)	50.7	73.3	70.0	72.7	100.0
Gemma-3 (12B)	63.3	84.0	72.7	76.0	100.0
Qwen-2.5 (32B)	69.1	87.2	87.9	75.2	99.3
GPT-5-mini	91.9	98.7	95.3	97.3	99.3
Haiku-4.5 (Claude)	92.0	96.7	95.3	97.3	100.0

cessing and the OCR-first approach. Some models improve on many, if not all, **REST** benchmarks (e.g., GPT-5-mini, Phi-4 and Gemma-3-4B). However, the performance is also hurt substantially in some cases (e.g., InternVL3 2B on SOEBENCH). These results reinforce that OCR is not the confounding factor behind cross-model inconsistency, because the OCR-first approach does not yield consistent results and can actually hurt performance.

5. Benchmarking Results: REST+

We evaluate the same 15 MLLMs on our more challenging **REST+** benchmark, which contains image permutations with different fonts, resolution and colours. We evaluate the full set of ten image permutations per question and show

Table 4. **‘OCR first, then solve’ strategy can hurt performance.** We show the difference in accuracy (Δ) when MLLMs are instructed to first transcribe text from images before solving the task (OCR-correct subset). This does not yield consistent improvements across models and benchmarks. Full table in Appendix.

Model	MMLU		A12-ARC		GSM-Sym		SoEBench	
	Img	Δ	Img	Δ	Img	Δ	Img	Δ
Deepseek-Small	51.2	-13.1	70.0	-12.8	60.9	-3.4	4.7	-1.3
GPT-4o-mini	77.0	-5.0	89.9	-2.1	89.0	-0.4	62.0	+2.7
GPT-5-mini	87.8	+1.2	93.3	+1.1	94.1	+0.8	97.3	+0.7
Gemini-2.5 FL	78.4	-0.3	88.8	+0.2	50.5	+9.1	49.3	-5.3
Gemma-3-12B	72.2	-0.0	88.6	+1.0	88.1	+1.5	76.0	-5.3
Haiku-4.5	85.1	+0.1	92.7	-0.7	94.1	-1.3	97.3	-1.3
InternVL3-14B	83.0	-0.1	94.2	0.0	93.0	-0.5	72.7	-1.3
Phi-4	31.4	+2.2	35.0	+9.0	54.5	+6.2	17.3	-4.0
Qwen-2.5-32B	83.5	-3.5	92.9	-2.5	93.0	+0.4	75.2	+10.1

the results in Table 5. First, OCR performance is worse than for REST, as expected, because of the lower image resolutions. Second, the overall RER and CFR scores are lower for **REST+** than the basic version. The high CFR values (best: 30.2%) indicate that the input format matters: a different font, colour, or resolution can determine whether an answer is correct or incorrect. Finally, the model rankings differ from the OCR correct and the full set of questions. In particular, one model (InternVL3-14B) achieves the highest consistency scores on both, but others (Claude Haiku and Mistral) decrease significantly when considering the full set of questions.

5.1. RQ3a: Impact of resolution on inconsistency

We investigate the effect of different resolutions on inconsistency, see Table 6 for the best-performing models. Some models maintain consistent performance across different resolutions (InternVL3-14B and Gemini). However, oth-

Table 5. **REST+ scores for 15 MLLMs show varying degrees of cross-modal inconsistency.** RER and CFR scores are given for all questions and for those where models obtain correctly recognised the rendered text (OCR \checkmark).

Model	REST+ (OCR \checkmark)		REST+ (All Questions)		
	RER \uparrow	CFR \downarrow	RER \uparrow	CFR \downarrow	OCR \uparrow
Deepseek-Tiny	5.8	96.2	3.5	97.0	82.7
Phi-4	11.3	89.2	7.6	90.5	80.1
Phi-3.5	14.1	91.7	7.6	93.8	77.3
Deepseek-Small	24.5	76.9	21.0	78.9	91.5
InternVL3 (2B)	31.8	69.9	27.5	72.0	85.2
Gemma-3 (4B)	36.8	67.0	27.6	73.9	70.5
Gemma-3 (12B)	53.2	46.7	49.9	49.5	86.6
Qwen-2.5 (7B)	53.9	45.8	49.7	49.7	91.7
GPT-4o-mini	61.2	38.8	60.8	39.0	95.7
Mistral-Small	61.6	38.4	55.4	44.2	89.6
Gemini-2.5 Flash Lite	63.8	32.4	62.7	33.3	95.9
Haiku-4.5 (Claude)	65.7	34.0	45.8	54.0	82.6
GPT-5-mini	67.6	32.4	64.8	35.1	94.5
Qwen-2.5 (32B)	69.4	30.4	65.5	33.7	90.7
InternVL3 (14B)	72.1	27.9	69.3	30.2	88.8

Table 6. **Cross-modal consistency varies across font resolution.** RER and OCR across DPI levels show that some models maintain consistent performance, while others do not (set of all questions).

Model	DPI@50		DPI@100		DPI@200	
	RER \uparrow	OCR \uparrow	RER	OCR	RER	OCR
Haiku-4.5 (Claude)	53.6	57.4	74.4	92.2	76.3	94.3
Qwen-2.5 (7B)	58.1	79.1	66.5	97.0	69.8	97.2
Gemma-3 (12B)	62.5	73.9	64.9	90.9	63.6	92.9
Mistral-Small	60.0	73.9	76.9	95.8	78.4	96.6
GPT-4o-mini	70.8	93.5	73.2	96.4	72.2	96.9
Gemini-2.5 F Lite	70.6	95.1	70.4	96.2	70.9	96.3
GPT-5-mini	71.5	89.6	82.7	95.9	83.6	97.1
Qwen-2.5 (32B)	71.2	77.8	78.1	95.5	79.5	96.9
InternVL3 (14B)	75.3	88.9	79.0	89.1	79.9	88.6

ers show decreased OCR and RER performance at DPI@50 (Claude and Gemma), indicating that OCR capabilities at this resolution are a key factor in inconsistency (if not controlling for OCR performance). Therefore, we conclude that OCR capabilities should be considered when evaluating cross-modal inconsistency (e.g., Claude Haiku would have shown worse performance if we had not controlled for incorrect OCR).

Comparing efficiency of text and vision tokens Different image resolutions result in different numbers of visual tokens for many models. Wei et al. [40] have developed DeepSeek-OCR to maintain performance when compressing text into fewer visual tokens. We investigate whether current MLLMs could already obtain similar efficiency gains (where we use efficiency to describe obtaining similar or better performance with fewer tokens).

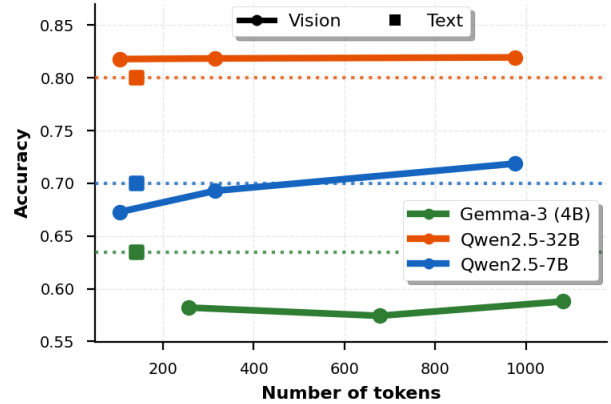


Figure 3. **Models generally achieve higher text accuracy despite using fewer text tokens.** Current MLLMs need more vision tokens than text tokens to achieve the same accuracy, except for Qwen2.5-VL-32B, where fewer vision tokens obtain higher accuracy (OCR-correct subset).

Figure 3 shows text and image accuracy for three models at different DPI levels (50, 100, 200). Qwen2.5-32B demonstrates higher efficiency for vision, where less image tokens achieve 1.8% higher accuracy than text tokens (104 image tokens on average ($\sigma = 24$) at 50 DPI, vs. 142 text tokens). In contrast, Qwen2.5-7B shows less visual token efficiency at DPI@50, where text and vision accuracies differ significantly despite comparable token counts. In fact, only Qwen-32B exhibits the property of being more efficient with vision tokens while scoring better than with text tokens, all other models (see Appendix) either perform better with text or are more efficient with text. Note that image token counts exclude text instruction tokens. Interestingly, InternVL3-14B (the most consistent model on REST+) uses approximately 1600 visual tokens (regardless of DPI), while requiring only 160 text tokens on average, a 10:1 ratio. Although InternVL3 shows higher image accuracy than text accuracy, text tokens are more efficient. Since we lack exact token usage data for closed-source models, we cannot perform the same analysis. Nevertheless, Claude, the GPTs, and Gemini all obtain better performance for text than image accuracy at DPI@50.

DeepSeek-OCR presents an interesting direction for future research in making MLLMs more efficient. However, we observe that current models reason substantially better through the text modality and that fewer text tokens than vision tokens are needed to obtain the same performance (except for Qwen2.5-VL-32B). In other words, although models can correctly read text from images, they do not necessarily reason as well as they would with conventional text input.

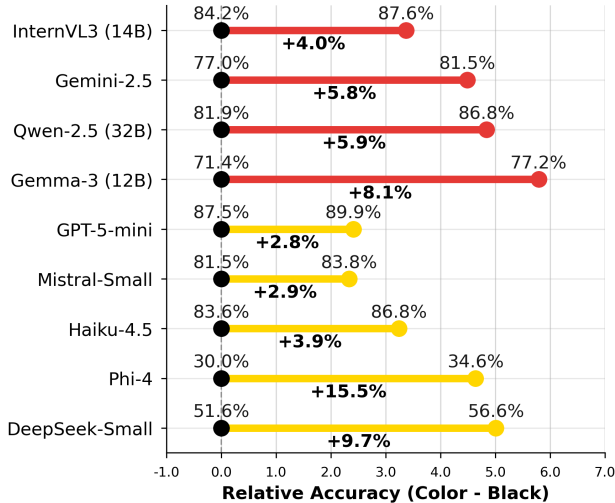


Figure 4. **Colored text makes models perform better.** Relative improvements from either red or yellow text compared to black (OCR-correct subset).

5.2. RQ3b: Impact of font, colour on inconsistency

Surprisingly, font families show no clear differences in image accuracy, despite our initial expectations that cursive fonts would be harder to read. Most models stay within 2% absolute difference between fonts (only Phi-3.5 shows a 5.3% difference between DejaVu Sans and Courier New). Unexpectedly, most models perform better with coloured text than black text (Figure 4, DPI@200). Most MLLMs achieve higher accuracies with red or yellow fonts. Multiple models even obtain more than 5% relative improvement (e.g. DeepSeek-Small and Qwen-32B). We show more detailed results in the Appendix.

6. RQ4: Multimodal Representation Analysis

To answer **RQ4**, we aim to find the internal mechanisms that determine a model’s inconsistency. We postulate that models with higher similarities between their modalities score better on our benchmark. We calculate the *implicit alignment score* introduced by Shukor and Cord [34]. They define alignment in terms of cosine similarity: the higher the similarity score, the more the representations point in the same direction. They find that this score is a proxy metric for task performance on several multimodal benchmarks.

We use Imagenet [12], containing natural images and text labels. We expand the data by generating images containing written-down labels (see Figure 5). This simulates our **REST** benchmark which contains text-only, image-only and mixed data. We perform the experiments on 1,000 samples. Note that we exclude our **REST** benchmark, because our aim is to assess the correlation between our benchmark and the modality gap (**RQ4**). To this end, we analyse real-

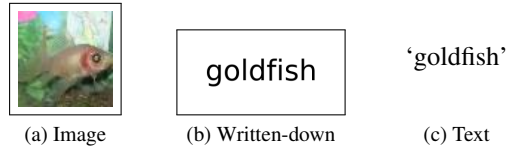


Figure 5. Samples from our 3-type Imagenet categories dataset for evaluating the intermediate representations of MLLMs.

world images which are expected to be more familiar to MLLMs and yield in-distribution image representations.

Since we need to retrieve the hidden activations, we focus on the open-source MLLMs from our previous experiments. We use three prompts, depending on the modality (1) images: “The main object in this image can be described as:”, (2) written-down images: “Repeat the text on the image:”, and (3) text: “Repeat the text {text_prompt}:”.

These prompts are chosen such that the model is prompted to (1) give the same response in all modalities and (2) pay attention to the image or text of interest. For our analysis, we represent each sequence of tokens by computing its mean embedding. Given image tokens $\mathbf{I} = \{\mathbf{i}_1, \mathbf{i}_2, \dots, \mathbf{i}_A\} \in \mathbb{R}^{A \times d}$ where A denotes the number of image tokens and d is the embedding dimension, we compute their mean representation $\bar{\mathbf{i}} = \frac{1}{A} \sum_{k=1}^A \mathbf{i}_k \in \mathbb{R}^d$. Similarly, for text tokens $\mathbf{T} = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_B\} \in \mathbb{R}^{B \times d}$ where B denotes the number of text tokens, we compute $\bar{\mathbf{t}} = \frac{1}{B} \sum_{k=1}^B \mathbf{t}_k \in \mathbb{R}^d$. We measure the similarity between image and text representations using cosine similarity:

$$\text{sim}(\mathbf{I}, \mathbf{T}) = \frac{\bar{\mathbf{i}} \cdot \bar{\mathbf{t}}}{\|\bar{\mathbf{i}}\| \|\bar{\mathbf{t}}\|}. \quad (4)$$

To evaluate cross-modal alignment, we compute a retrieval accuracy score that measures how often each sample correctly retrieves its corresponding pair from the opposite modality. For a dataset containing N paired image-text samples, we determine for each image i which text j has the highest similarity:

$$y_i = \text{argmax}_{j \in \{1, \dots, N\}} \text{sim}(\mathbf{I}_i, \mathbf{T}_j). \quad (5)$$

The similarity score is then computed as the fraction of correct retrievals:

$$\text{Similarity score} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}[y_i = i], \quad (6)$$

where $\mathbb{1}[\cdot]$ is the indicator function that equals 1 when the retrieved text index matches the correct pairing and 0 otherwise. This metric effectively measures the accuracy of cross-modal retrieval, where higher scores indicate better alignment between the image and text representations. We compute this similarity score at each layer of each model

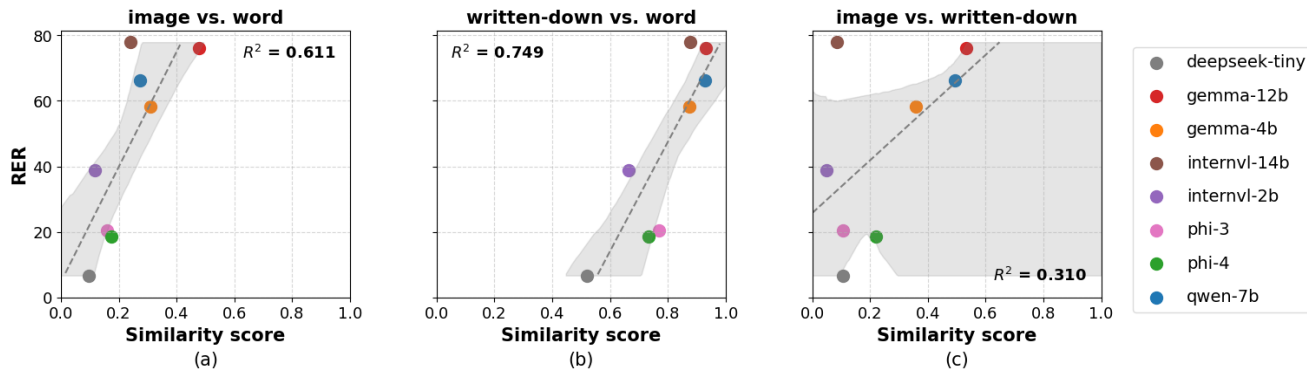


Figure 6. **Benchmark performance is correlated to the similarity of modalities.** The similarity between image vs. word (a) and written-down vs. word (b) representations correlates with RER (as determined using our **REST** benchmark). R^2 denotes the variance explained by the fitted line, and the grey area shows the bootstrapped 95% confidence interval.

and report the maximum score across all layers. We repeat this process bidirectionally (image-to-text and text-to-image retrieval), yielding multiple similarity scores per model that characterize the cross-modal alignment from different perspectives. Figure 6 shows that similarity scores are correlated with the Render-Equivalence Rate (RER), especially for the similarity between images and words and between written-down labels and words. This indicates that matching questions are very similar to each other in consistent models (i.e. models with high RER). Note that the similarity scores between written-down labels and words are higher overall, indicating that written-down representations are more similar to matching word representations than image representations are.

7. Related Work

Most MLLM benchmarks are visual question-answering tasks [14, 17, 23, 24, 36, 44] and do not include data samples where the same question is presented across different modalities. The Omni-R benchmark [8] does contain paired questions in text, image, video, and audio, based on MMLU-Pro [39]. However, it evaluates inconsistency only in closed-source MLLMs and does not control for OCR effects. MMMU-Pro [45] contains a vision-only subtask in which the text is rendered alongside the original visuals. Our work differs in making the OCR task as simple as possible, whereas MMMU-Pro renders text in screenshots or natural photos. The vision-language consistency dataset of Zhang et al. [46] includes multiple existing text benchmarks, but evaluates only GPT-4V [3] and similarly does not control for OCR. MMIR [42] focuses on a different form of multimodal inconsistency: detecting semantic mismatches between text and vision. Alonso et al. [5] show that MLLMs are unable to consistently match the same entity across modalities, which contain complementary information, and not identical (like ours). Similarly, Samson et al. [32] find cases where MLLMs handle the concept of pri-

vacy differently in text and vision, and Sim et al. [35] study modality collapse (ignoring certain modalities) and find that the text modality often dominates.

Current literature agrees on the existence of a modality gap in different MLLMs, where text and image embeddings occupy different regions in the joint embedding space [20, 30, 33, 34]. Liang et al. show the modality gap for the first time for CLIP, which they attribute to the two modalities initialising a narrow embedding cone, which is preserved by the loss function [20]. This phenomenon has since been documented more broadly across MLLMs [34]. Shukor and Cord actually find a positive correlation between the cosine similarity of the text and image representations within one prompt and task performance [34]. Eslami and de Melo find that mitigating this gap in CLIP (by parameter sharing) improves the performance on downstream tasks [13].

8. Discussion and Conclusion

In this work, we present the **REST** and **REST+** benchmarks to evaluate cross-modal inconsistency in MLLMs. Through our experiments, we find that there is considerable cross-modal inconsistency in state-of-the-art MLLMs, which answers **RQ1**. This varies across models, with cross-modal inconsistency scores ranging from 10% to 90%, depending on the MLLM. We find that GPT-5-mini [29] and Claude Haiku 4.5 [6] achieve the highest consistency scores and that almost all MLLMs perform best in the text modality.

Despite only evaluating on questions with correct OCR, we find that the cross-modal inconsistency persists. This indicates that cross-modal inconsistency is not simply a by-product of poor OCR performance, which answers **RQ2**. We find that neither rendering images as text (OCR first) nor rendering text as images solves the inconsistency. We find that font resolution has an effect on cross-modal inconsistency, where higher DPI correlates with more consistent MLLM performance across modalities, even when controlling for OCR performance. This answers **RQ3a**. In other

words, although models can correctly read text from images, they do not necessarily reason as well as they would with conventional text input. We show that font does not have a clear impact on cross-modal inconsistency, while text colour does (**RQ3b**). In our multimodal representation analysis experiment, we examine the internal representations of the same concept in different modalities. We observe a positive correlation between the cross-modal similarity of internal representations and our RER consistency score. This implies that the direction of representations from matching samples is linked to modality inconsistency, which answers **RQ4**. Future work could study whether this effect is causal, in particular by optimising the representations to be more similar and observing whether the RER consistency score increases.

References

- [1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL <https://arxiv.org/abs/2404.14219>, 2: 6, 2024. 4
- [2] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 4
- [3] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1, 8
- [4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1
- [5] Iñigo Alonso, Gorka Azkune, Ander Salaberria, Jeremy Barnes, and Oier Lopez de Lacalle. Vision-language models struggle to align entities across modalities. *arXiv preprint arXiv:2503.03854*, 2025. 1, 8
- [6] Anthropic. Claude haiku 4.5. <https://www.anthropic.com>, 2025. Large language model by Anthropic. 3, 4, 8
- [7] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 4
- [8] Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, Yandong Li, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, et al. Omnixr: Evaluating omni-modality language models on reasoning across modalities. *arXiv preprint arXiv:2410.12219*, 2024. 1, 4, 8
- [9] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. 1, 4, 7
- [10] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018. 2
- [11] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 1
- [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 7
- [13] Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in clip, 2024. 1, 8
- [14] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, 2017. 8
- [15] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 2, 3
- [16] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020. 1, 4, 7
- [17] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 8
- [18] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 4
- [19] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 2
- [20] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 1, 8
- [21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1
- [22] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

- [23] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 1, 8
- [24] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *ICLR*, 2024. <https://arxiv.org/abs/2310.02255>. 8
- [25] Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, 2022. 1
- [26] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209, 2021. 1
- [27] Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024. 1, 2, 4, 7
- [28] Mistral AI. Mistral-small-3.1-24b-instruct. <https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503>, 2025. Version 2503. 4
- [29] OpenAI. Gpt-5 mini. <https://openai.com>, 2025. Compact large language model by OpenAI. 3, 4, 8
- [30] Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham M. Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces. In *Second Conference on Language Modeling*, 2025. 1, 8
- [31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1
- [32] Laurens Samson, Nimrod Barazani, Sennay Ghebreab, and Yuki M Asano. Privacy-aware visual language models. *arXiv e-prints*, pages arXiv-2405, 2024. 1, 8
- [33] Dong Shu, Haiyan Zhao, Jingyu Hu, Weiru Liu, Ali Payani, Lu Cheng, and Mengnan Du. Large vision-language model alignment and misalignment: A survey through the lens of explainability. *arXiv preprint arXiv:2501.01346*, 2025. 1, 8
- [34] Mustafa Shukor and Matthieu Cord. Implicit multimodal alignment: On the generalization of frozen llms to multimodal inputs. *Advances in Neural Information Processing Systems*, 37:130848–130886, 2024. 1, 7, 8
- [35] Mong Yuan Sim, Wei Emma Zhang, Xiang Dai, and Biaoyan Fang. Can vlms actually see and read? a survey on modality collapse in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24452–24470, 2025. 1, 8
- [36] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, 2019. 8
- [37] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3, 4
- [38] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 4
- [39] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024. 8
- [40] Haoran Wei, Yaofeng Sun, and Yukun Li. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*, 2025. 1, 6
- [41] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 4
- [42] Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. *ACL*, 2025. <https://arxiv.org/abs/2502.16033>. 8
- [43] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024. 1
- [44] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *CVPR*, 2024. 1, 8, 13
- [45] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 15134–15186, 2025. 1, 8
- [46] Xiang Zhang, Senyu Li, Ning Shi, Bradley Hauer, Zijun Wu, Grzegorz Kondrak, Muhammad Abdul-Mageed, and Laks V. S. Lakshmanan. Cross-modal consistency in multimodal large language models, 2024. 1, 3, 8
- [47] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 4

Same Content, Different Answers: Cross-Modal Inconsistency in MLLMs

Supplementary Material

9. Benchmark Implementation Details

We provide complete specifications for reproducing the **REST** benchmark experiments, including prompt templates and dataset examples. All code, data, and model outputs will be released publicly upon acceptance to facilitate future research on cross-modal consistency.

9.1. REST

Images In Figure 7, we show examples of both mixed and image modalities for randomly selected questions from MMLU, ARC, and GSM8K-Symbolic. Samples for SOEBENCH are provided in the following section. We render all images at DPI 200 on a white background using black DejaVu Sans font to maximise readability and ensure consistent OCR performance. To minimise OCR complexity, we exclude questions exceeding 800 characters and those containing LaTeX formatting, as mathematical notation can introduce ambiguity in text recognition tasks. The mixed modality format depends on the benchmark structure: for multiple-choice questions (MMLU, ARC), the answer options appear as text while the question context is rendered as an image; for open-ended tasks (GSM8K-Symbolic, SOEBENCH), the problem context appears as an image while the final question is presented as text.

Prompts Figure 8 presents the prompt templates for the four core tasks in the **REST** benchmark. We encourage Chain-of-Thought reasoning across all modalities and maintain consistent instructions, only adjusting for how a model should solve a question. For the OCR task, we explicitly instruct models to perform only text transcription without solving the problem, as some models would otherwise attempt to provide answers alongside the transcription.

9.2. REST+

For the more challenging benchmark, we create 10 image permutations per question (Figure 9). We generate 9 combinations using three font families (DejaVu Sans, Courier New, and Cursive) and three resolutions (50, 100, and 200 DPI). We specifically chose these distinct font types to evaluate how cross-modal consistency varies across different typographic styles. Additionally, we create one color variant using DejaVu Sans at 200 DPI, with colors assigned cyclically using modulo operation from a set of six standard colors (red, green, blue, cyan, magenta, and yellow). We manually verified that text remains legible at 50 DPI for all font families to ensure that performance differences reflect reasoning capabilities rather than readability issues.

We limit our evaluation to these 10 permutations for computational feasibility, as each question requires 21 forward passes (1 text, 10 OCR verification, and 10 image-based solving). Questions are sampled from MMLU at 10% per subject area to maintain subject balance across the reduced dataset. With 1,085 questions and 21 forward passes per question, we perform 22,785 evaluations per MLLM. We exclude the mixed modality from REST+ as it would require 10 additional forward passes per question, substantially increasing computational costs.

For the prompts, we use the same templates as in **REST** depicted in Figure 8, maintaining consistency in evaluation instructions across both benchmarks.

9.3. SOEBENCH

To ensure zero data contamination and minimise OCR complexity, we introduced SOEBENCH, a novel system-of-equations benchmark specifically designed for evaluating cross-modal consistency. Each puzzle presents n letter variables (A through E for $n \in \{3, 4, 5\}$) that must be solved to find integer values between 1 and 9. The benchmark consists of 150 puzzles total, with 50 puzzles for each value of n .

Each puzzle contains n clue equations plus one final equation to solve. Variables appear with coefficients ranging from 1 to 3, combined using basic arithmetic operations (+, -, *). For example, a clue equation might read “ $2A + B = 15$ ” while the final equation presents “ $3B - A + 2C = ?$ ” where the model must determine the result. We ensure each puzzle has a unique solution by verifying that only one assignment of values satisfies all clue equations simultaneously.

For the mixed modality, we render the clue equations as an image while presenting the final equation as text. For the image modality, all equations, including the final question, are rendered together. We use DejaVu Sans font at 200 DPI on a white background with black text to maximise readability. The restricted symbol set (digits 0-9, letters A-E, and basic operators) ensures that OCR performance remains near-perfect for most models, allowing us to isolate reasoning capabilities from recognition challenges.

Figure 10 shows example puzzles with varying numbers of variables.

Prompts Figure 11 presents the prompt templates. While the text, image, and mixed prompts follow similar Chain-of-Thought structures as **REST**, we introduce a delimiter-based output format using to clearly separate reasoning from the final answer, as models extensively reason.

Which statement correctly describes a property of a type of matter?

(a) ARC - Mixed modality

Which statement correctly describes a property of a type of matter?
(A) Air is a mixture of gases.
(B) Ice is a mixture of gases.
(C) Air is a liquid.
(D) Ice is a liquid.

(b) ARC — Image modality

A group of 54 students has various hobbies. 35 like to dance, 2 like to play badminton, and the rest like to either paint or bake.

(c) GSM8K — Mixed modality

A group of 54 students has various hobbies. 35 like to dance, 2 like to play badminton, and the rest like to either paint or bake. How many like to paint if the number that like to bake is twice the number that prefer playing badminton?

(d) GSM8K — Image modality

In men, specimens for gonococcal cultures are most commonly obtained from which of the following structures?

(e) MMLU — Mixed modality

In men, specimens for gonococcal cultures are most commonly obtained from which of the following structures?
(A) Anus
(B) Bladder
(C) Urethra
(D) Testicle

(f) MMLU — Image modality

Figure 7. Examples of ARC, GSM8K, and MMLU questions in mixed and image modalities from our REST benchmark. In the mixed modality, part of the content (e.g., multiple-choice options or context) is provided as text while the rest is rendered as an image. In the image modality, the entire content is rendered as a single image.

For the OCR verification task, we provide explicit formatting instructions requiring models to number each equation sequentially (1), (2), (3), etc. This structured format serves two purposes: it allows us to verify that models correctly identify the semantic structure of the equations, and it simplifies parsing and validation of OCR outputs, as we noticed different (correct) output formats for models.

9.4. Model Configuration and Hardware

We conduct all experiments using vLLM [19] for computational efficiency. Temperature is set to 0 for deterministic outputs across all models where configurable. For proprietary models, we apply the following settings: GPT-5-mini uses minimal reasoning effort (temperature control unavailable), Claude Haiku-4.5 has thinking mode disabled, and Gemini-2.5 Flash Lite operates with thinking budget set to 0. Despite these computational optimisations, all models receive identical Chain-of-Thought prompting instructions to ensure fair comparison.

Open-source models follow vLLM’s recommended configurations for optimal performance. Experiments run on single-GPU systems: NVIDIA RTX 6000 Ada (48GB VRAM) for most models, and NVIDIA H100 (80GB VRAM) for larger models (Mistral-Small-3.1-24B and Qwen2.5-32B) due to memory requirements.

Solve the following question.

Think step by step, but put the answer (A, B, C or D) on the very last line, preceded by `Answer :`. Do not write anything else on that line.

Example:
Reasoning...
Answer: A

Question

(a) Text modality

Solve the question in the image.

Think step by step, but put the answer (A, B, C or D) on the very last line, preceded by `Answer :`. Do not write anything else on that line.

Example:
Reasoning...
Answer: C

(b) Image modality

Read the question in the image and choose from the options below.

Think step by step, but put the answer (A, B, C or D) on the very last line, preceded by `Answer :`. Do not write anything else on that line.

Example:
Reasoning...
Answer: B

Multiple Choice Options

(c) OCR verification

You are given an image that contains text. You must do the following:

1. Do not solve the question; just transcribe the text exactly as it appears.
2. Do not add extra commentary, only transcribe. Please transcribe now.

Figure 8. Prompt templates used in the **REST** benchmark for evaluating cross-modal consistency across MMLU, ARC, GSM-Symbolic. Each modality (text, image, mixed) receives task-specific instructions while maintaining consistent Chain-of-Thought reasoning requirements and standardized answer formatting. The OCR verification prompt (d) ensures that text recognition capabilities are assessed independently from reasoning performance. For non-MMLU benchmarks, prompts follow identical structures with minor adaptations: mixed modality varies based on whether questions are multiple-choice (options as text) or open-ended (questions as text, context as image).

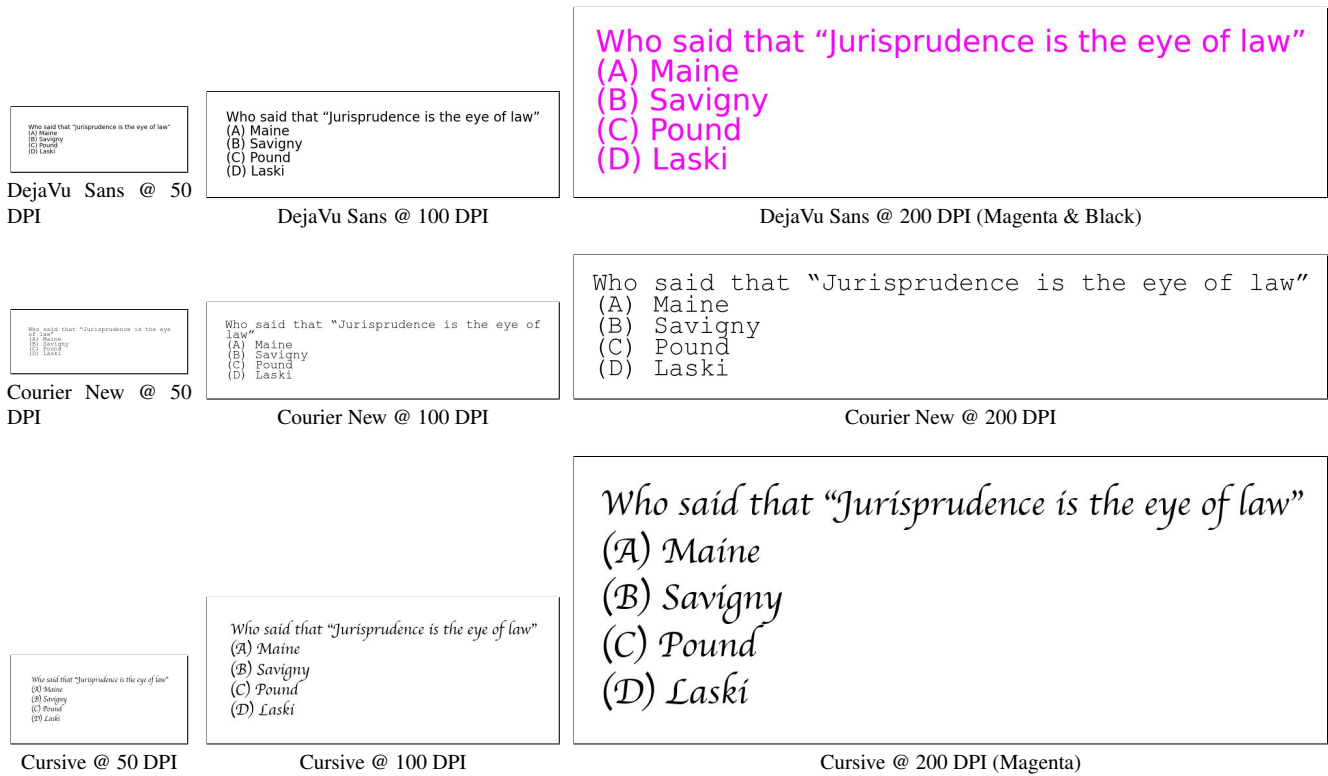


Figure 9. Visual permutations in REST+ benchmark showing the same MMLU question rendered with varying resolutions (columns: 50, 100, 200 DPI) and fonts (rows: DejaVu Sans, Courier New, Cursive). We manually verified that images are legible at DPI 50.

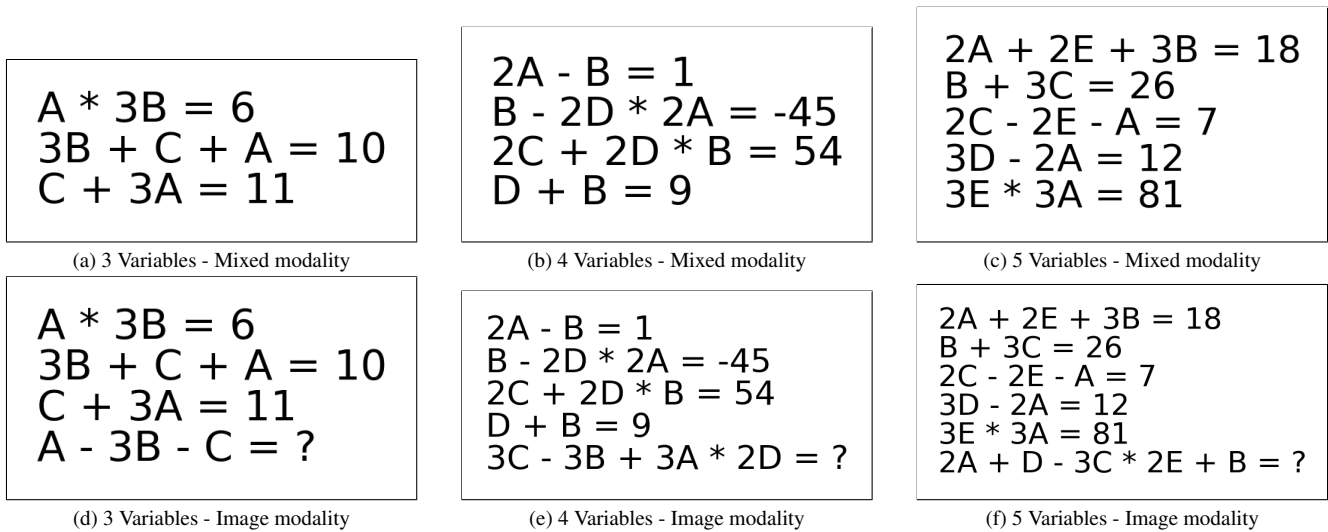


Figure 10. SOEBENCH examples with increasing complexity from 3 to 5 variables. Top row: mixed images containing only clue equations, used for the mixed modality where the final equation is presented as text. Bottom row: images including all equations with the final equation used for the image modality. Each puzzle requires finding integer values (1-9) for letter variables that satisfy all equations simultaneously. **Font rendering is consistent across all images (DejaVu Sans, 200 DPI); perceived size differences are due to automatic figure scaling for layout consistency.**

This puzzle contains equations with letters representing natural numbers from 1 to 9:

[All equations]

Provide the outcome of the final equation.

Think step by step. When completely finished, output:

####

Then write only the final answer:

Example:

Reasoning step 1 ...

Reasoning step 2 ...

####

Answer: 28

(a) Text modality

This puzzle contains equations with letters representing natural numbers from 1 to 9.

Solve the puzzle in the image, provide the outcome of the final equation.

Think step by step. When completely finished, output:

####

Then write only the final answer:

Example:

Reasoning step 1 ...

Reasoning step 2 ...

####

Answer: 28

(b) Image modality

This puzzle contains equations with letters representing natural numbers from 1 to 9:

Read the information in the image and then solve the equation below:

[Final equation with ?]

Think step by step. When completely finished, output:

####

Then write only the final answer:

Example:

Reasoning step 1 ...

Reasoning step 2 ...

####

Answer: 28

(c) Mixed modality

You are given an image that contains a list of separate mathematical equations.

1. Do not solve or simplify these equations; just transcribe them exactly as they appear.
2. Retain the same order and use the following numbering, (1), (2), (3) per equation.
3. List 1 equation per item, for example (1) $3a + 2b + c = 11$
4. Put each equation on its own line.
5. Use plain text as output, the operations that you can use are '*', '+', '-' and '='.
6. Do not add extra commentary-only transcribe equations. Format your output like so:
(1) $2a + 3b = 10$
(2) $a + 3c = 30$
(3) $2b + 5c = ?$
Please transcribe now.

(d) OCR verification

Figure 11. Prompt templates used for SOEBENCH evaluation. Each modality receives specific instructions for solving systems of equations with letter variables. For the mixed modality, clue equations are provided as images while the final equation appears as text. For OCR, we instruct the models for a specific format, as models generate different types of correct output formats.

10. Extended Results

This section presents comprehensive results for the **REST** benchmarks, including performance metrics across all evaluation conditions and detailed breakdowns by modality.

10.1. REST

Cross-Modal Consistency Analysis Tables 7 and 8 present RER and CFR scores for the OCR-correct subset and the complete set, respectively. Results are given per benchmark (MMLU, ARC, GSM8k-Symbolic, SOEBENCH) with corresponding OCR accuracy rates. Notably, DeepSeek-Tiny achieves 0% OCR accuracy on SOEBENCH, resulting in worst-case consistency scores for this model-benchmark combination. The minimal difference between OCR-correct and full dataset scores validates our approach of constraining OCR complexity: cross-modal inconsistency persists even when text recognition is perfect, confirming that the phenomenon stems from fundamental reasoning differences rather than recognition failures.

Modality-Specific Performance Tables 9 and 10 provide accuracy scores across text, image, and mixed modalities for both evaluation subsets. Character Error Rate (CER) metrics provide a quantitative assessment of OCR performance across models. We additionally report results for the “OCR-first” strategy, which yields mixed outcomes, improving performance for some model-benchmark combinations while degrading others.

Notably, Max Modal Coverage (MMC) analysis reveals that several models achieve near-perfect solvability when considering their best-performing modality: GPT-5-mini attains 100% MMC on SOEBENCH, 96.0% on ARC, and 97.3% on GSM8k-Symbolic. However, this high coverage does not translate to consistent performance. The same model exhibits variation across modalities, highlighting the gap between potential and realised performance under different input formats.

Distribution of Solvable Questions Figures 12 and 13 present staircase plots visualizing the cumulative distribution of correctly solved questions across modality combinations. These plots reveal distinct patterns: models with high consistency (e.g., GPT-5-mini) show steep initial rises with most questions solved across all modalities, while less consistent models exhibit gradual staircases indicating substantial modality-dependent performance variation.

Table 7. **REST Consistency performance across VLMs when OCR Correct.** Columns show Render-Equivalence Rate (RER), Cross-Modality Failure Rate (CFR), and perfect OCR rate (OCR) for each benchmark component. On the left, scores for **REST**, which are the mean over all RER and CFR scores. **Results include only questions with correct OCR.**

Model	REST (Avg.)		MMLU [16]			AI2-ARC [9]			GSM-Sym [27]			SoEBench		
	RER	CFR	RER	CFR	OCR	RER	CFR	OCR	RER	CFR	OCR	RER	CFR	OCR
Deepseek-Tiny	6.6	98.0	2.3	97.8	100.0	4.0	95.6	100.0	20.3	98.7	100.0	0.0	100.0	100.0
Phi-4	14.9	82.3	8.8	89.8	100.0	12.1	86.7	100.0	37.5	57.3	100.0	1.3	95.5	100.0
Phi-3.5	19.4	79.3	21.5	80.3	100.0	28.4	70.7	100.0	27.7	66.2	100.0	0.0	100.0	100.0
InternVL3 (2B)	32.9	63.7	46.1	52.6	100.0	62.0	37.0	100.0	19.0	76.4	100.0	4.7	88.9	100.0
Deepseek-Small	35.9	60.9	42.6	56.9	100.0	57.8	40.6	100.0	42.7	50.1	100.0	0.7	95.8	100.0
Gemma-3 (4B)	53.9	42.3	53.3	46.1	100.0	65.0	33.8	100.0	55.4	38.7	100.0	41.9	50.5	100.0
Gemini-2.5-flash-lite	54.3	40.3	74.9	19.5	100.0	87.6	8.6	100.0	41.4	51.2	100.0	13.3	81.8	100.0
Qwen-2.5 (7B)	64.6	31.7	73.8	25.0	100.0	86.2	13.8	100.0	73.8	22.7	100.0	24.7	65.4	100.0
GPT-4o-mini	71.3	26.0	74.5	24.6	100.0	89.1	10.8	100.0	82.4	15.1	100.0	39.3	53.6	100.0
Mistral-Small	73.6	23.9	75.8	23.0	100.0	88.8	10.7	100.0	88.5	9.9	100.0	41.3	52.0	100.0
Gemma-3 (12B)	75.8	21.3	69.8	28.6	100.0	87.7	11.4	100.0	82.4	14.7	100.0	63.3	30.7	100.0
InternVL3 (14B)	78.4	19.6	81.7	17.8	100.0	93.8	5.6	100.0	87.5	11.5	100.0	50.7	43.6	100.0
Qwen-2.5 (32B)	84.7	13.6	83.3	15.1	100.0	92.9	5.3	100.0	93.4	5.8	100.0	69.1	28.0	100.0
Haiku-4.5 (Claude)	90.3	8.9	84.5	14.9	100.0	92.2	6.5	100.0	92.6	6.9	100.0	92.0	7.4	100.0
GPT-5-mini	90.7	8.7	85.8	13.3	100.0	93.3	6.4	100.0	91.9	7.0	100.0	91.9	8.1	100.0

Table 8. **REST Consistency performance across VLMs on all questions.** Columns show Render-Equivalence Rate (RER), Cross-Modality Failure Rate (CFR), and perfect OCR rate (OCR) for each benchmark component. On the left, scores for **REST**, which are the mean over all RER and CFR scores. **Results include all questions (including OCR incorrect).**

Model	REST (Avg.)		MMLU [16]			AI2-ARC [9]			GSM-Sym [27]			SoEBench		
	RER	CFR	RER	CFR	OCR	RER	CFR	OCR	RER	CFR	OCR	RER	CFR	OCR
Deepseek-Tiny	6.5	98.1	2.2	97.9	90.6	3.9	95.8	96.8	19.9	98.7	93.6	0.0	100.0	0.0
Phi-4	14.5	82.8	7.8	90.9	84.9	12.1	86.9	97.5	36.7	58.0	93.5	1.3	95.5	100.0
Phi-3.5	19.3	79.5	22.0	80.4	79.2	28.2	70.9	91.4	26.9	66.7	90.5	0.0	100.0	100.0
InternVL3 (2B)	32.6	63.9	45.3	53.5	88.6	61.6	37.2	96.6	18.7	76.1	77.5	4.7	88.9	100.0
Deepseek-Small	35.9	60.9	42.4	57.1	95.3	57.8	40.6	99.3	42.6	50.1	96.9	0.7	95.8	100.0
Gemma-3 (4B)	52.3	44.0	53.6	47.3	73.4	63.5	35.1	90.2	53.4	40.3	84.2	38.7	53.2	86.0
Gemini-2.5-flash-lite	54.1	40.4	74.4	19.8	96.9	87.7	8.5	99.6	41.1	51.2	96.9	13.3	81.8	100.0
Qwen-2.5 (7B)	64.3	31.9	73.4	25.3	97.9	86.2	13.8	99.8	72.9	22.9	97.5	24.7	65.4	100.0
GPT-4o-mini	71.2	26.1	74.2	24.8	97.5	89.0	10.9	99.6	82.2	15.3	98.2	39.3	53.6	100.0
Mistral-Small	73.4	24.1	75.6	23.3	97.7	88.8	10.7	99.9	88.1	10.4	95.9	41.3	52.0	100.0
Gemma-3 (12B)	75.5	21.6	69.5	28.9	93.6	87.7	11.4	98.9	81.7	15.4	93.5	63.3	30.7	100.0
InternVL3 (14B)	78.1	19.9	81.1	18.2	90.6	93.8	5.5	95.5	86.9	12.1	95.2	50.7	43.6	100.0
Qwen-2.5 (32B)	84.5	13.7	83.2	15.3	96.7	92.9	5.3	99.5	93.1	6.1	94.5	68.7	28.0	99.3
Haiku-4.5 (Claude)	90.1	9.1	84.3	15.0	97.7	92.2	6.4	99.7	91.8	7.8	95.6	92.0	7.4	100.0
GPT-5-mini	90.7	8.8	85.7	13.4	98.5	93.3	6.4	99.7	91.6	7.2	98.4	92.0	8.0	99.3

Table 9. **Modality-specific accuracies (Text, Image, Mixed) across REST benchmarks.** MMC indicates maximum modal coverage (percentage of questions solved in at least one modality). The ‘OCR First’ metrics represent the accuracy when first prompting to OCR and then solving the question. **Results include only questions with correct OCR.**

MMLU							GSM8K-Symbolic						
Model	Text	Mixed	Image	OCR First	CER	MMC	Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	82.9	81.4	83.0	83.0	0.0	90.0	InternVL3 (14B)	91.2	91.9	93.0	92.5	0.0	96.8
InternVL3 (2B)	61.6	59.0	56.6	55.4	0.0	79.6	InternVL3 (2B)	52.1	40.3	53.1	64.6	0.0	76.8
Mistral-Small	84.1	78.9	82.5	82.6	0.0	91.3	Mistral-Small	92.6	91.7	93.3	93.2	0.0	96.5
Phi-3.5	61.1	41.5	32.5	39.3	0.0	75.6	Phi-3.5	66.9	42.9	48.3	45.2	0.0	78.3
Phi-4	47.0	44.2	31.4	33.6	0.0	80.1	Phi-4	76.5	55.8	54.5	60.7	0.0	85.9
Qwen-2.5 (32B)	83.3	84.1	83.5	80.0	0.0	90.0	Qwen-2.5 (32B)	93.6	93.1	93.0	93.4	0.0	95.7
Qwen-2.5 (7B)	72.6	71.8	72.9	72.9	0.0	82.4	Qwen-2.5 (7B)	84.4	82.0	82.1	83.0	0.0	91.9
Haiku-4.5 (Claude)	90.1	87.3	85.1	85.2	0.0	93.7	Haiku-4.5 (Claude)	95.6	94.9	94.1	92.8	0.0	97.7
Deepseek-Small	54.9	54.9	51.2	38.1	0.0	75.2	Deepseek-Small	67.3	54.7	60.9	57.5	0.0	80.4
Deepseek-Tiny	29.5	26.6	25.3	17.6	0.0	68.7	Deepseek-Tiny	15.5	0.8	0.7	0.3	0.0	16.0
Gemini-2.5-flash-lite	81.6	79.7	78.4	78.2	0.0	88.1	Gemini-2.5-flash-lite	79.4	60.1	50.5	59.5	0.0	84.8
Gemma-3 (12B)	77.6	76.7	72.2	72.2	0.0	87.1	Gemma-3 (12B)	91.3	87.7	88.1	89.5	0.0	95.1
Gemma-3 (4B)	68.0	64.7	57.2	57.5	0.0	81.3	Gemma-3 (4B)	82.2	70.5	67.4	70.0	0.0	89.6
GPT-4o-mini	84.4	77.2	77.0	71.9	0.0	89.8	GPT-4o-mini	91.4	86.7	89.0	88.6	0.0	95.2
GPT-5-mini	89.3	86.5	87.8	89.0	0.0	93.4	GPT-5-mini	95.1	93.7	94.1	95.0	0.0	97.4

SOEBENCH							AI2-ARC						
Model	Text	Mixed	Image	OCR First	CER	MMC	Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	73.3	70.0	72.7	71.3	0.0	88.7	InternVL3 (14B)	92.9	92.9	94.2	94.2	0.0	95.7
InternVL3 (2B)	14.0	14.7	33.3	9.3	0.0	42.0	InternVL3 (2B)	76.0	74.0	74.0	72.8	0.0	89.7
Mistral-Small	60.7	62.0	62.0	63.3	0.0	82.0	Mistral-Small	92.7	90.0	92.6	92.0	0.0	95.8
Phi-3.5	1.3	1.3	0.7	0.7	0.0	3.3	Phi-3.5	73.3	54.3	37.9	57.0	0.0	83.6
Phi-4	13.3	13.3	17.3	13.3	0.0	29.3	Phi-4	59.9	48.5	35.0	44.1	0.0	86.1
Qwen-2.5 (32B)	87.2	87.9	75.2	85.2	0.0	96.0	Qwen-2.5 (32B)	92.9	92.9	92.9	90.4	0.0	95.2
Qwen-2.5 (7B)	48.0	46.0	48.7	36.7	0.0	71.3	Qwen-2.5 (7B)	86.1	87.0	88.3	88.9	0.0	93.1
Haiku-4.5 (Claude)	96.7	95.3	97.3	96.0	0.0	99.3	Haiku-4.5 (Claude)	93.3	93.3	92.7	92.0	0.0	95.7
Deepseek-Small	8.0	7.3	4.7	3.3	0.0	16.0	Deepseek-Small	69.1	68.7	70.0	57.2	0.0	85.3
Deepseek-Tiny	0.7	0.0	0.0	0.0	70.5	0.7	Deepseek-Tiny	34.2	28.7	27.5	19.7	0.0	70.6
Gemini-2.5-flash-lite	42.7	34.7	49.3	44.0	0.0	73.3	Gemini-2.5-flash-lite	90.4	89.0	88.8	89.0	0.0	93.2
Gemma-3 (12B)	84.0	72.7	76.0	70.7	0.0	91.3	Gemma-3 (12B)	90.5	90.2	88.6	89.6	0.0	94.4
Gemma-3 (4B)	75.2	58.9	57.4	55.8	0.0	84.5	Gemma-3 (4B)	79.2	78.2	71.4	72.7	0.0	90.5
GPT-4o-mini	65.3	61.3	62.0	64.7	0.0	83.3	GPT-4o-mini	93.6	89.3	89.9	87.8	0.0	95.4
GPT-5-mini	98.7	95.3	97.3	98.0	0.0	100.0	GPT-5-mini	94.6	92.7	93.3	94.4	0.0	96.0

Table 10. **Modality-specific accuracies (Text, Image, Mixed) across REST benchmarks.** MMC indicates maximum modal coverage (percentage of questions solved in at least one modality). The ‘OCR First’ metrics represent the accuracy when first prompting to OCR and then solving the question. **Results include all questions (including OCR incorrect).**

MMLU							GSM8K-Symbolic						
Model	Text	Mixed	Image	OCR First	CER	MMC	Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	82.4	81.0	82.5	82.6	0.5	89.7	InternVL3 (14B)	90.6	91.4	92.6	92.2	0.1	96.6
InternVL3 (2B)	61.2	58.5	56.0	55.0	1.1	79.4	InternVL3 (2B)	50.8	38.7	51.8	62.2	3.8	74.8
Mistral-Small	83.9	78.7	82.3	82.3	0.0	91.3	Mistral-Small	92.2	91.3	92.9	92.8	0.0	96.3
Phi-3.5	58.1	40.3	31.3	37.1	4.5	72.8	Phi-3.5	65.5	41.9	46.8	43.7	0.7	76.9
Phi-4	45.0	43.2	30.6	32.5	0.6	79.1	Phi-4	75.5	55.1	53.7	60.1	0.1	85.1
Qwen-2.5 (32B)	83.3	84.1	83.3	79.7	0.1	90.0	Qwen-2.5 (32B)	93.2	92.8	92.7	92.5	0.1	95.4
Qwen-2.5 (7B)	72.4	71.5	72.7	72.7	0.0	82.3	Qwen-2.5 (7B)	83.6	81.2	81.3	82.1	0.0	91.1
Haiku-4.5 (Claude)	90.0	87.2	85.1	85.2	0.1	93.7	Haiku-4.5 (Claude)	95.4	94.4	93.6	92.4	0.1	97.7
Deepseek-Small	54.8	54.6	51.1	37.7	0.2	75.1	Deepseek-Small	67.2	54.6	60.7	57.2	0.1	80.2
Deepseek-Tiny	29.5	26.2	25.2	17.5	1.3	68.6	Deepseek-Tiny	15.2	0.8	0.7	0.3	0.4	15.8
Gemini-2.5-flash-lite	81.3	79.4	78.0	77.6	0.1	87.9	Gemini-2.5-flash-lite	78.8	59.6	50.1	59.2	0.0	84.2
Gemma-3 (12B)	77.4	76.5	71.9	71.8	0.2	87.0	Gemma-3 (12B)	90.9	87.2	87.7	89.0	0.3	95.0
Gemma-3 (4B)	64.0	60.2	53.5	53.8	3.6	76.8	Gemma-3 (4B)	81.1	68.1	65.4	68.2	1.8	88.4
GPT-4o-mini	84.2	77.0	76.7	71.7	0.0	89.6	GPT-4o-mini	91.5	86.6	88.8	88.3	0.0	95.3
GPT-5-mini	89.2	86.5	87.7	89.0	0.0	93.4	GPT-5-mini	94.9	93.5	93.8	94.5	0.0	97.3

SOEBENCH							AI2-ARC						
Model	Text	Mixed	Image	OCR First	CER	MMC	Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	73.3	70.0	72.7	71.3	0.0	88.7	InternVL3 (14B)	93.0	93.2	94.3	94.3	0.2	95.8
InternVL3 (2B)	14.0	14.7	33.3	9.3	0.0	42.0	InternVL3 (2B)	75.8	73.6	73.5	72.3	1.7	89.5
Mistral-Small	60.7	62.0	62.0	63.3	0.0	82.0	Mistral-Small	92.7	90.0	92.6	92.0	0.0	95.8
Phi-3.5	1.3	1.3	0.7	0.7	0.0	3.3	Phi-3.5	74.0	54.7	37.6	55.5	5.3	84.2
Phi-4	13.3	13.3	17.3	13.3	0.0	29.3	Phi-4	59.9	48.3	35.0	44.0	0.1	86.0
Qwen-2.5 (32B)	86.7	87.3	74.7	84.7	0.0	95.3	Qwen-2.5 (32B)	93.0	93.0	92.9	90.4	0.0	95.2
Qwen-2.5 (7B)	48.0	46.0	48.7	36.7	0.0	71.3	Qwen-2.5 (7B)	86.1	87.1	88.3	88.9	0.0	93.1
Haiku-4.5 (Claude)	96.7	95.3	97.3	96.0	0.0	99.3	Haiku-4.5 (Claude)	93.3	93.3	92.7	92.0	0.0	95.7
Deepseek-Small	8.0	7.3	4.7	3.3	0.0	16.0	Deepseek-Small	69.1	68.8	70.0	57.2	0.0	85.4
Deepseek-Tiny	0.7	0.0	0.0	0.0	70.5	0.7	Deepseek-Tiny	34.6	28.9	27.1	19.6	0.5	70.6
Gemini-2.5-flash-lite	42.7	34.7	49.3	44.0	0.0	73.3	Gemini-2.5-flash-lite	90.4	89.0	88.9	89.0	0.2	93.2
Gemma-3 (12B)	84.0	72.7	76.0	70.7	0.0	91.3	Gemma-3 (12B)	90.6	90.2	88.6	89.6	0.0	94.4
Gemma-3 (4B)	74.0	55.3	52.7	52.0	1.2	82.7	Gemma-3 (4B)	77.9	76.7	70.4	71.2	1.4	89.6
GPT-4o-mini	65.3	61.3	62.0	64.7	0.0	83.3	GPT-4o-mini	93.6	89.3	90.0	87.7	0.0	95.4
GPT-5-mini	98.7	95.3	97.3	98.0	0.2	100.0	GPT-5-mini	94.6	92.7	93.3	94.4	0.0	96.0

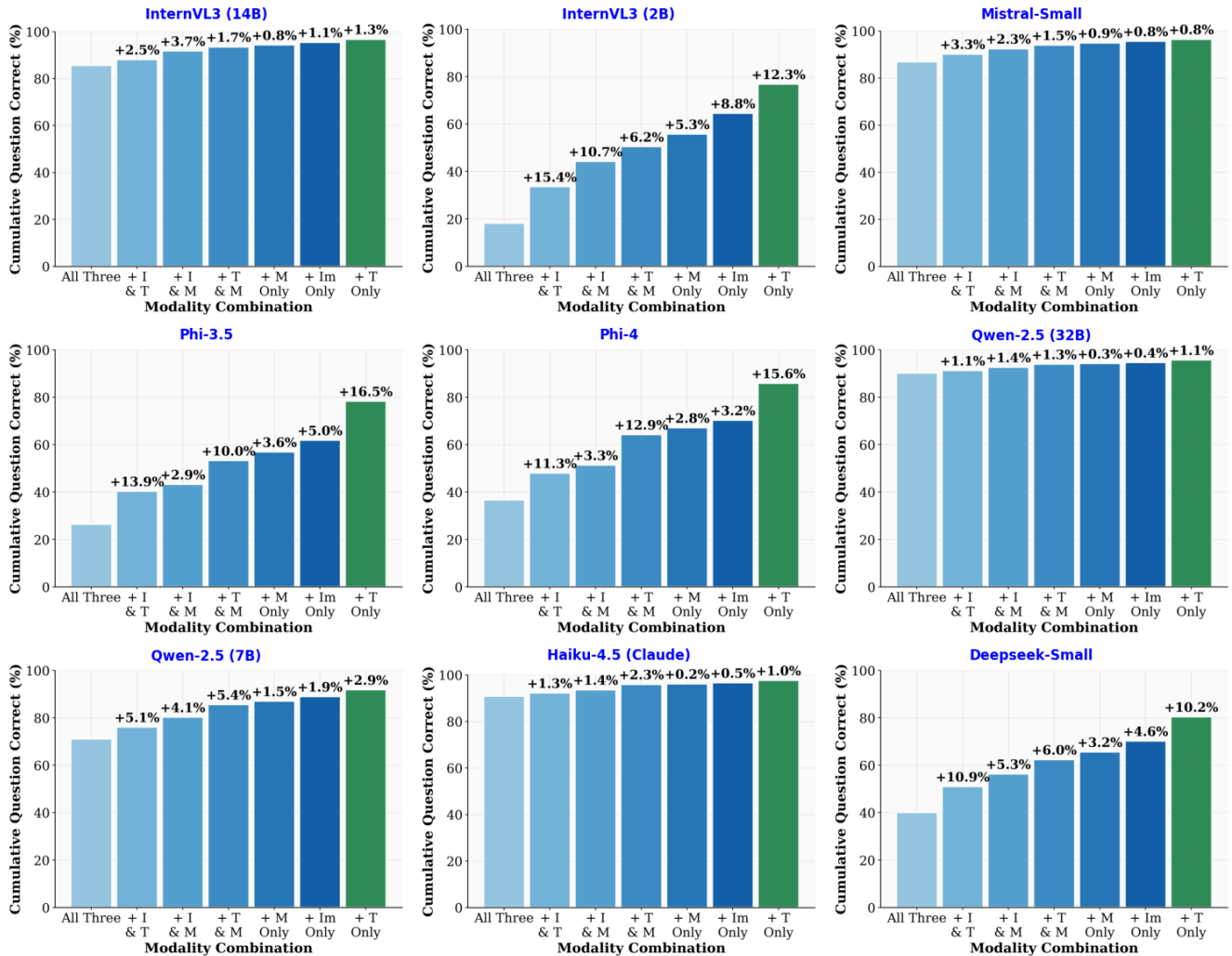


Figure 12. Cumulative distribution of correctly solved questions across modality combinations for models 1-8. Each step represents questions solvable in progressively fewer modalities, with the leftmost portion showing questions solved consistently across all three modalities.

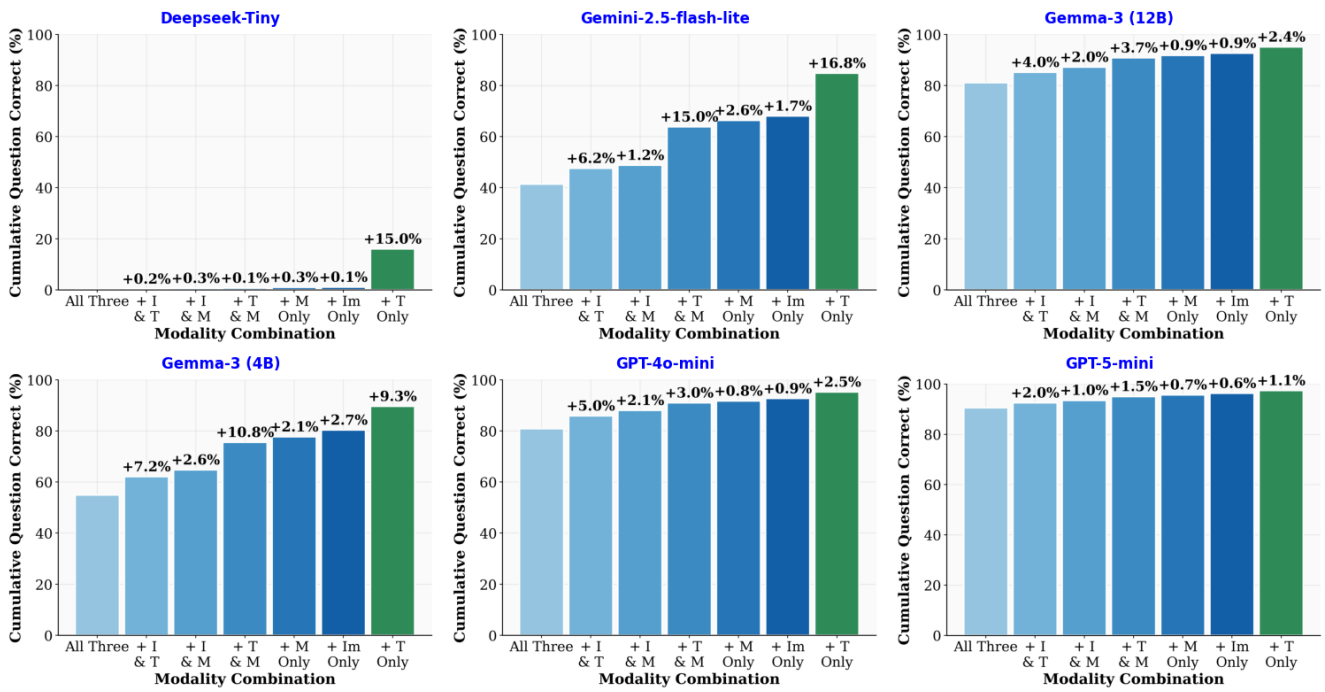


Figure 13. Cumulative distribution of correctly solved questions across modality combinations for models 9-15. Models with higher cross-modal consistency show larger proportions of questions solved in all modalities (leftmost region).

10.2. REST+

Cross-modal Consistency Tables 11 and 12 present **REST+** performance metrics across all visual permutations. We report RER consistency scores, modality-specific accuracies, and OCR success rates for both the OCR-correct subset and the complete dataset. Image accuracy represents the mean performance across all 10 visual permutations per question. Consistent with **REST** findings, text modality systematically outperforms image modality across all models. The reduced OCR accuracy demonstrates that lower resolution substantially impacts both text recognition and downstream reasoning.

Table 11. **REST+ performance on OCR-correct subset.** Render Equivalence Rate (RER) and modality-specific accuracies for questions where text recognition was perfect. Image accuracy is averaged across all visual permutations (3 fonts \times 3 resolutions + colour variant) if OCR was correct.

Model	RER	Text	Image
Deepseek-Tiny	5.8	28.6	24.4
Phi-4	11.3	43.9	32.7
Phi-3.5	14.1	57.6	32.6
Deepseek-Small	24.5	55.5	51.9
InternVL3 (2B)	31.8	62.9	57.3
Gemma-3 (4B)	36.8	63.5	58.2
Gemma-3 (12B)	53.2	77.2	73.3
Qwen-2.5 (7B)	53.9	70.0	69.8
GPT-4o-mini	61.2	83.9	77.6
Mistral-Small	61.6	81.9	80.9
Gemini-2.5-flash-lite	63.8	82.7	77.3
Haiku-4.5 (Claude)	65.7	85.5	83.7
GPT-5-mini	67.6	89.4	86.3
Qwen-2.5 (32B)	69.4	80.0	82.0
InternVL3 (14B)	72.1	82.1	82.6

Influence of DPI Tables 13 and 14 present resolution-dependent performance analysis for **REST+**, showing image accuracy and RER consistency scores across three DPI levels (50, 100, 200). Results reveal distinct model robustness patterns: while some models (DeepSeek-Small, Mistral-Small) exhibit performance degradation at lower resolutions, others (Qwen2.5-32B, InternVL3-14B) maintain stable performance across the different DPI levels.

Influence of Font Family Tables 15 and 16 present font-specific performance metrics across DejaVu Sans, Courier New, and Cursive font families. Most models demonstrate consistent performance across font families. Claude Haiku-4.5 shows a performance drop specifically for Courier New in the complete dataset (but not in the OCR-correct subset), suggesting that it struggles with OCR on monospace font characteristics rather than having inherent reasoning limitations.

Table 12. **REST+ performance on complete dataset.** Comprehensive evaluation, including all questions regardless of OCR success. Lower RER scores compared to the OCR-correct subset reflect the compounding effects of recognition errors and reasoning inconsistencies.

Model	RER	Text	Image	OCR
Deepseek-Tiny	3.5	28.6	24.8	82.7
Phi-3.5	7.6	57.6	31.0	77.3
Phi-4	7.6	43.9	32.4	80.1
Deepseek-Small	21.0	55.5	51.0	91.5
InternVL3 (2B)	27.5	62.9	55.5	85.2
Gemma-3 (4B)	27.6	63.5	53.5	70.5
Haiku-4.5 (Claude)	45.8	85.5	80.5	82.6
Qwen-2.5 (7B)	49.7	70.0	68.9	91.7
Gemma-3 (12B)	49.9	77.2	71.7	86.6
Mistral-Small	55.4	81.9	79.5	89.6
GPT-4o-mini	60.8	83.9	77.4	95.7
Gemini-2.5-flash-lite	62.7	82.7	77.0	95.9
GPT-5-mini	64.8	89.4	85.9	94.5
Qwen-2.5 (32B)	65.5	80.0	81.3	90.7
InternVL3 (14B)	69.3	82.1	82.0	88.8

Table 13. **Image performance on REST+ for different DPI levels(OCR-correct subset).** Image accuracy and RER consistency scores stratified by resolution (50, 100, 200 DPI) for questions with perfect text recognition.

Model	Image Accuracy			RER		
	50	100	200	50	100	200
Deepseek-Small	51.6	52.4	51.6	41.0	40.0	41.4
Deepseek-Tiny	23.9	24.7	24.3	14.9	12.8	10.0
GPT-4o-mini	78.0	78.7	76.1	72.6	73.7	72.8
GPT-5-mini	84.0	87.4	87.5	75.1	83.4	84.0
Gemini-2.5-flash-lite	78.1	77.3	77.0	71.5	71.0	71.7
Gemma-3 (12B)	76.1	73.4	71.4	73.0	67.3	65.7
Gemma-3 (4B)	58.2	57.4	58.8	59.3	54.4	54.6
Haiku-4.5 (Claude)	82.1	84.4	83.6	78.5	77.3	76.8
InternVL3 (14B)	80.7	82.6	84.2	78.4	81.5	82.2
InternVL3 (2B)	56.4	57.8	58.2	45.7	46.2	46.9
Mistral-Small	78.3	82.0	81.5	69.7	77.8	79.0
Phi-3.5	32.1	32.6	32.6	26.7	26.8	25.8
Phi-4	34.4	34.8	30.0	21.7	18.0	14.1
Qwen-2.5 (32B)	81.8	81.8	81.9	77.9	78.7	79.8
Qwen-2.5 (7B)	67.3	69.3	71.9	65.8	67.4	71.4

Influence of colour Similarly, Tables 17 and 18 analyse the effect of text colour on image modality accuracy. We report performance for black text across all resolutions and specifically at 200 DPI to enable fair comparison with coloured variants (all rendered at 200 DPI). Remarkably, every evaluated model achieves higher accuracy with at least one colour compared to black text at equivalent resolution.

Token Usage Finally, we show the number of tokens used by MLLMs in text and at the different DPI levels in Table 19. As mentioned in the paper, we see that fewer text tokens are needed to achieve the same level of accuracy, or more

Table 14. **Image performance on REST+ for different DPI levels** (complete set of questions). Image accuracy, RER consistency and OCR correct scores stratified by resolution (50, 100, 200 DPI) for all questions.

Model	Img Acc.			RER			OCR		
	50	100	200	50	100	200	50	100	200
Deepseek-Small	49.7	51.8	51.2	32.7	36.7	39.2	85.3	93.4	94.9
Deepseek-Tiny	24.5	24.7	24.6	6.6	7.3	6.0	72.1	85.8	88.4
GPT-4o-mini	77.6	78.6	76.1	70.8	73.2	72.2	93.5	96.4	96.9
GPT-5-mini	83.1	87.1	87.5	71.5	82.7	83.6	89.6	95.9	97.1
Gemini-2.5-FL	77.7	77.2	76.8	70.6	70.4	70.9	95.1	96.2	96.3
Gemma-3 (12B)	71.6	72.6	70.9	62.5	64.9	63.6	73.9	90.9	92.9
Gemma-3 (4B)	52.1	54.7	53.8	41.7	46.6	44.9	63.1	76.5	71.7
Haiku-4.5	73.4	84.0	83.8	53.6	74.4	76.3	57.4	92.2	94.3
InternVL3 (14B)	80.2	82.1	83.5	75.3	79.0	79.9	88.9	89.1	88.6
InternVL3 (2B)	54.2	55.8	56.7	39.7	40.4	41.4	84.5	84.7	86.1
Mistral-Small	74.6	81.8	81.4	60.0	76.9	78.4	73.9	95.8	96.6
Phi-3.5	30.2	30.9	31.4	18.0	16.5	17.1	75.7	76.0	79.6
Phi-4	32.8	34.6	30.3	10.9	13.3	10.2	62.7	86.3	88.5
Qwen-2.5 (32B)	79.7	81.7	81.9	71.2	78.1	79.5	77.8	95.5	96.9
Qwen-2.5 (7B)	65.2	69.2	71.5	58.1	66.5	69.8	79.1	97.0	97.2

Table 15. **Image performance on REST+ for different font families (OCR-correct subset)**. Image accuracy and RER consistency scores for different fonts for questions with perfect text recognition.

Model	Image Accuracy			RER		
	font	Deja. Sans	Cour. New	Curs.	Deja. Sans	Cour. New
Deepseek-Small	53.6	49.3	52.6	42.8	40.2	42.3
Deepseek-Tiny	24.5	23.6	24.8	10.5	13.0	11.8
GPT-4o-mini	77.3	77.3	78.2	71.4	72.1	73.7
GPT-5-mini	86.9	85.2	87.0	81.1	76.4	82.0
Gemini-2.5-flash-lite	78.6	76.6	77.1	72.0	70.8	71.5
Gemma-3 (12B)	73.6	72.8	74.0	66.7	68.4	67.4
Gemma-3 (4B)	57.5	58.4	58.5	54.2	58.6	55.2
Haiku-4.5 (Claude)	84.3	83.2	83.1	76.3	78.6	77.2
InternVL3 (14B)	82.4	82.4	82.6	81.1	80.4	80.3
InternVL3 (2B)	57.0	56.1	59.4	46.7	44.1	53.0
Mistral-Small	80.2	81.1	81.1	71.6	77.0	74.2
Phi-3.5	34.9	29.6	32.6	30.3	24.9	28.4
Phi-4	33.9	32.4	32.4	17.1	16.1	17.0
Qwen-2.5 (32B)	81.8	82.4	81.4	77.8	80.1	78.7
Qwen-2.5 (7B)	69.6	69.1	70.1	65.6	67.5	68.0

vision tokens are necessary to achieve higher accuracy than text, with the exception of Qwen2.5-VL-32B

Correlation of REST with common benchmarks We examine whether high-scoring MLLMs on general vision-language benchmarks show less cross-modal inconsistency. To this end, we plot the RER scores of **REST** and **REST+** against scores on MMMU [44], see Figure 14. The MMMU scores were found on public benchmarks¹. Generally, we

¹<https://mmmu-benchmark.github.io/#leaderboard>, <https://www.anthropic.com/news/claude-haiku-4-5>,

Table 16. **Image performance on REST+ for different font families (complete set of questions)**. Image accuracy and RER consistency scores for different fonts on all questions.

Model	Image Accuracy			RER		
	font	Deja. Sans	Cour. New	Curs.	Deja. Sans	Cour. New
Deepseek-Small	52.8	47.7	52.2	39.5	34.3	38.7
Deepseek-Tiny	24.9	23.8	25.1	6.2	6.2	6.9
GPT-4o-mini	77.1	77.2	77.9	70.6	71.5	72.6
GPT-5-mini	86.7	84.7	86.3	80.2	73.6	79.8
Gemini-2.5-flash-lite	78.4	76.5	76.8	71.4	70.1	70.8
Gemma-3 (12B)	72.1	70.6	72.4	63.0	61.3	63.1
Gemma-3 (4B)	53.6	53.2	53.8	44.0	43.0	44.6
Haiku-4.5 (Claude)	82.7	76.3	82.2	74.7	54.9	74.2
InternVL3 (14B)	82.0	81.7	82.0	78.8	77.1	78.5
InternVL3 (2B)	55.9	54.6	56.1	42.2	37.7	41.1
Mistral-Small	79.8	77.2	80.8	69.7	63.4	72.8
Phi-3.5	32.9	28.8	30.8	22.6	13.7	19.5
Phi-4	33.6	31.5	32.5	10.9	9.2	10.8
Qwen-2.5 (32B)	81.1	80.9	81.3	75.7	73.8	77.8
Qwen-2.5 (7B)	69.0	67.2	69.7	63.0	60.1	66.3

Table 17. **Text color effects on REST+ image accuracy (OCR-correct subset)**. Comparison of accuracy for black text at multiple resolutions (50, 100, 200 DPI) versus colored text variants (all at 200 DPI). Numbers indicate the percentage of correctly solved questions. **Results shown for OCR-correct subset only.**

Model	All DPI	DPI@200						
		Bl.	Bl.	R	G	B	Y	M
Deepseek-Small	51.9	51.6	55.5	52.3	51.7	56.6	49.7	49.1
Deepseek-Tiny	24.3	24.3	25.0	28.5	23.7	24.0	25.5	25.9
GPT-4o-mini	77.6	76.1	79.8	76.4	73.9	78.9	79.0	75.1
GPT-5-mini	86.3	87.5	84.9	87.6	80.4	89.9	86.4	86.8
Gemini-2.5 FL	77.5	77.0	81.5	75.0	71.0	74.7	74.4	77.4
Gemma-3 (12B)	73.5	71.4	77.2	70.2	70.9	67.1	72.6	75.6
Gemma-3 (4B)	58.1	58.8	62.0	58.7	61.7	56.0	58.1	53.3
Haiku-4.5 (Claude)	83.6	83.6	86.5	86.3	79.8	86.8	84.1	85.3
InternVL3 (14B)	82.5	84.2	87.6	83.2	81.0	84.6	84.1	79.5
InternVL3 (2B)	57.5	58.2	59.4	51.9	60.1	56.6	54.4	53.2
Mistral-Small	80.8	81.5	83.6	81.5	80.5	83.8	82.9	78.7
Phi-3.5	32.5	32.6	32.9	32.6	33.1	36.0	34.0	37.5
Phi-4	32.9	30.0	26.3	32.0	30.5	34.6	32.0	33.1
Qwen-2.5 (32B)	81.9	81.9	86.8	82.5	80.5	85.5	81.0	83.2
Qwen-2.5 (7B)	69.6	71.9	74.6	67.8	68.2	71.8	78.4	66.7

do find such a correlation, both for REST and REST+. Interestingly, the correlation seems low for models with high MMMU scores on REST+.

<https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct>, <https://huggingface.co/Qwen/Qwen2.5-VL-32B-Instruct>, <https://deepmind.google/models/gemini/flash-lite/>, <https://huggingface.co/OpenGVLab/InternVL3-2B>, <https://huggingface.co/microsoft/Phi-4-multimodal-instruct>, on Nov 20, 2025.

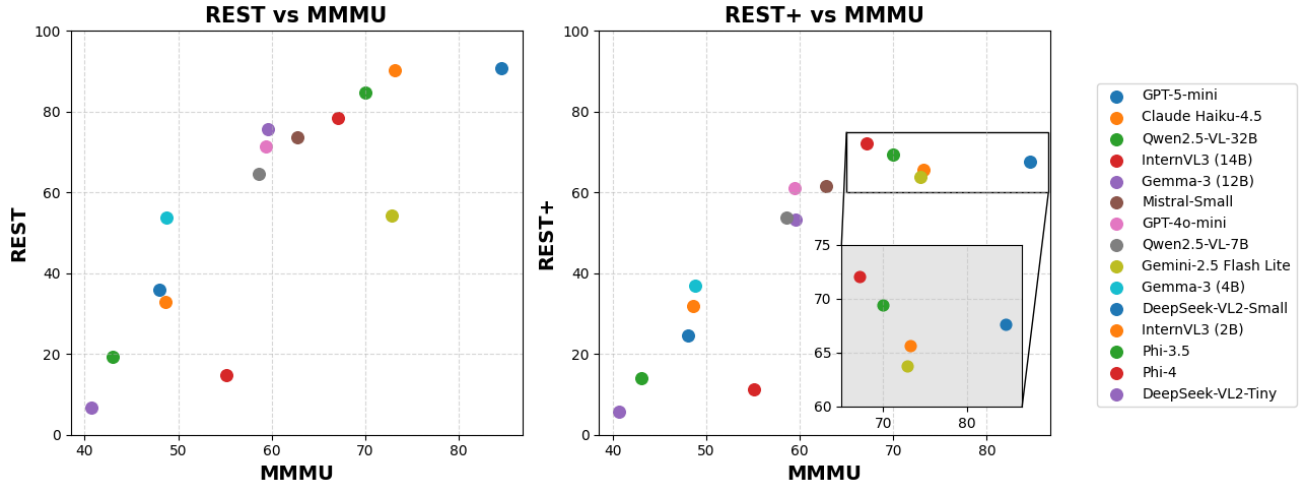


Figure 14. Models that perform well on MMMU also score well on REST and REST+. The zoom inset shows that models with high MMMU scores do not generally obtain high REST+ scores.

Table 18. **Text color effects on REST+ image accuracy (Complete set).** Comparison of accuracy for black text at multiple resolutions (50, 100, 200 DPI) versus colored text variants (all at 200 DPI). Numbers indicate the percentage of correctly solved questions. **Results shown for the complete dataset.**

Model	All DPI	DPI@200						
	Bl.	Bl.	R	G	B	Y	M	C
Deepseek-Small	50.9	51.2	54.7	51.4	51.9	56.9	49.4	48.3
Deepseek-Tiny	24.6	24.6	26.0	28.2	24.3	23.2	27.2	27.8
GPT-4o-mini	77.4	76.1	79.6	76.2	74.0	79.0	77.8	75.0
GPT-5-mini	85.9	87.5	85.1	87.3	80.7	90.1	86.7	86.7
Gemini-2.5 FL	77.2	76.8	81.8	74.0	71.3	73.5	73.9	77.8
Gemma-3 (12B)	71.7	70.9	76.2	70.7	70.2	68.5	71.7	75.0
Gemma-3 (4B)	53.5	53.8	57.5	50.8	54.1	50.8	56.1	52.2
Haiku-4.5 (Claude)	80.4	83.8	82.9	82.9	76.7	82.9	80.6	81.7
InternVL3 (14B)	81.9	83.5	86.2	82.9	80.1	84.5	84.4	80.6
InternVL3 (2B)	55.5	56.7	58.6	50.8	60.8	53.0	55.6	53.9
Mistral-Small	79.3	81.4	82.9	81.8	80.1	82.3	82.2	79.4
Phi-3.5	30.8	31.4	32.0	30.4	32.0	32.0	33.9	36.7
Phi-4	32.6	30.3	26.5	32.6	30.4	32.0	32.2	34.4
Qwen-2.5 (32B)	81.1	81.9	86.2	82.9	80.7	85.1	80.6	82.8
Qwen-2.5 (7B)	68.6	71.5	74.0	68.0	68.5	71.3	77.8	66.1

Table 19. **Token consumption across modalities in REST+.** Average number of tokens processed for text modality versus image modality at three resolutions (50, 100, 200 DPI). Vision token counts exclude instruction tokens and represent only image encoding.

Model	Text	Image		
		All.	DPI 50	DPI 100
Deepseek-Small	145	584	896	1668
Deepseek-Tiny	126	584	896	1668
GPT-4o-mini	-	-	-	-
GPT-5-mini	-	-	-	-
Gemini-2.5-flash-lite	-	-	-	-
Gemma-3 (12B)	141	256	698	1051
Gemma-3 (4B)	141	256	698	1051
Haiku-4.5 (Claude)	-	-	-	-
InternVL3 (14B)	168	1683	1616	1631
InternVL3 (2B)	168	1683	1616	1631
Mistral-Small	127	116	332	1011
Phi-3.5	154	611	542	577
Phi-4	122	425	706	1631
Qwen-2.5 (32B)	142	106	316	978
Qwen-2.5 (7B)	142	106	316	978