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

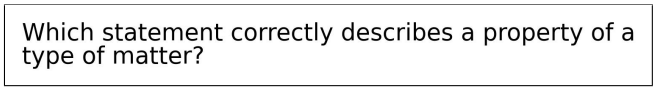
To ensure zero data contamination and minimise OCR complexity, we introduced SOEBENCH, a novel system-of-equations benchmark specifically designed for evaluating cross-modal consistency. Each puzzle presents n letter variables (A through E for $n \in \{3, 4, 5\}$) that must be solved to find integer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

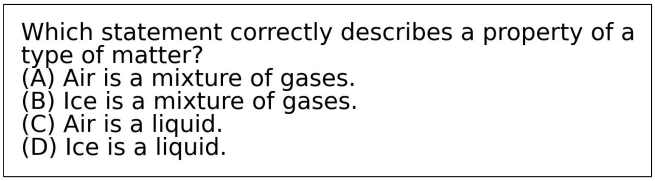
For the mixed modality, we render the clue equations as an image while presenting the final equation as text. For the image modality, all equations, including the final question, are rendered together. We use DejaVu Sans font at 200 DPI on a white background with black text to maximise readability. The 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.

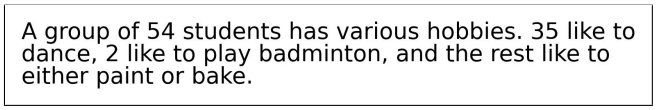
Prompts Figure 11 presents the prompt templates. While the text, image, and mixed prompts follow similar Chain-of-Thought structures as **REST**, we introduce a delimiter-based output format using to clearly separate reasoning from the final answer, as models extensively reason.



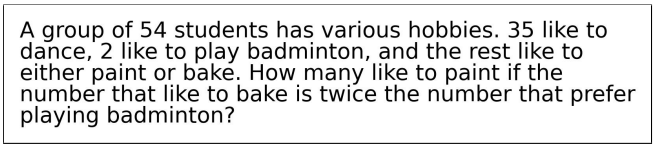
(a) ARC - Mixed modality



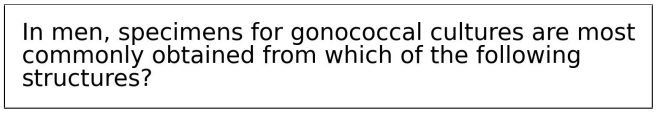
(b) ARC — Image modality



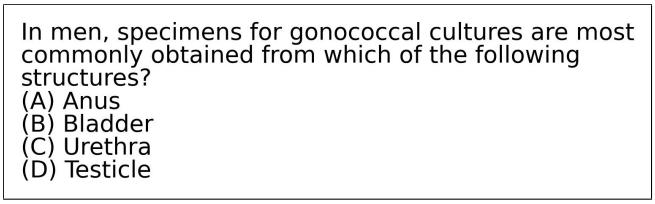
(c) GSM8K — Mixed modality



(d) GSM8K — Image modality



(e) MMLU — Mixed modality



(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 [17] 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

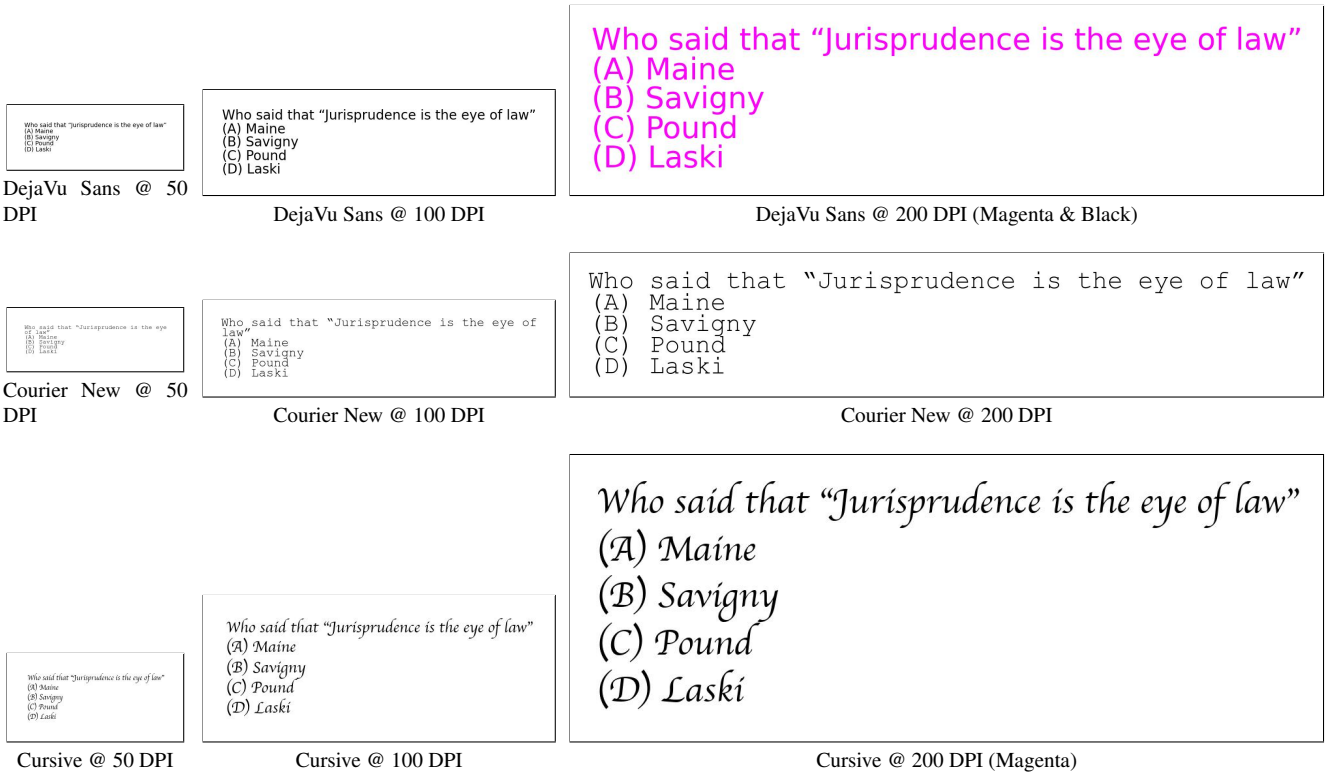
`Multiple Choice Options`

(c) Mixed modality

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.

(d) OCR verification

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).

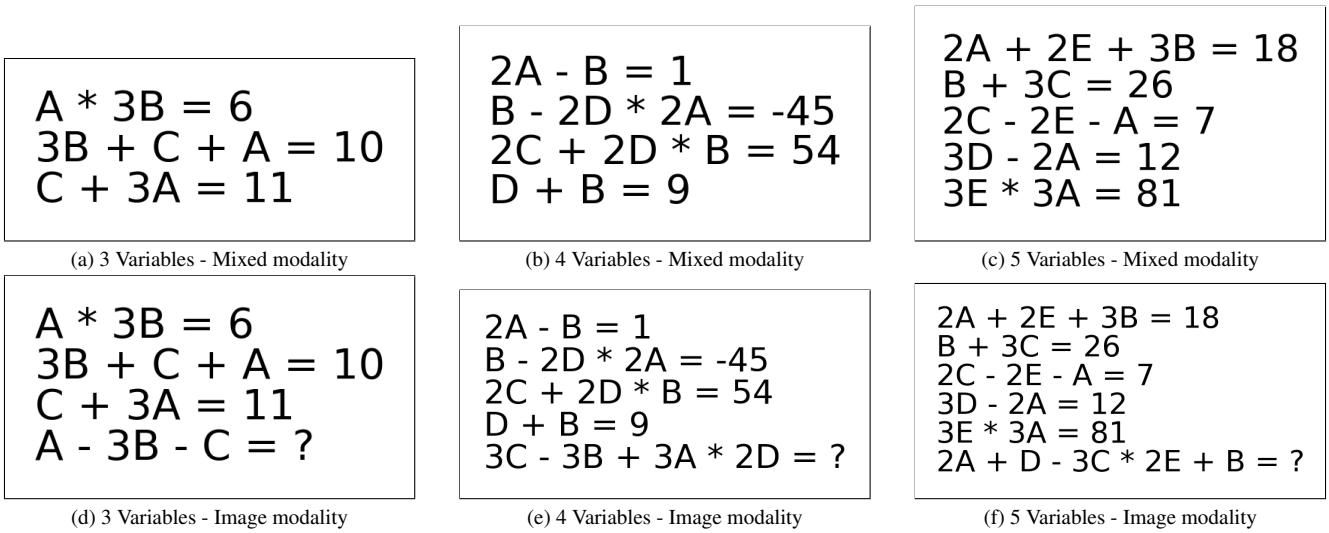


DejaVu Sans @ 50 DPI | DejaVu Sans @ 100 DPI | DejaVu Sans @ 200 DPI (Magenta & Black)

Courier New @ 50 DPI | Courier New @ 100 DPI | Courier New @ 200 DPI

Cursive @ 50 DPI | Cursive @ 100 DPI | Cursive @ 200 DPI (Magenta)

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.



(a) 3 Variables - Mixed modality | (b) 4 Variables - Mixed modality | (c) 5 Variables - Mixed modality

(d) 3 Variables - Image modality | (e) 4 Variables - Image modality | (f) 5 Variables - Image modality

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.

DeepSeek-OCR We also evaluate DeepSeek-OCR on our benchmarks. We observe that the model is not able to follow the instructions in our prompts. When we use their prescribed OCR prompt, $<image>$ *Free OCR*, we obtain a perfect OCR score on SOEBENCH. However, when using our OCR prompt, DeepSeek-OCR does not produce a single completely correct output. Moreover, for all reasoning questions, the model achieves a 0% score, indicating an inability to properly follow the instructions.

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.

Larger Models We evaluate InternVL-78B and Qwen-72B on REST (Table 11). Despite increased capacity, GPT-5-mini ($\sim$ 27-149B) still achieves the highest consistency scores, and inconsistency decreases only marginally; suggesting this is not purely a scaling issue. Due to resource constraints, we were not able to run bigger models on more data.

Mixed Design The mixed modality evaluates cross-modal integration, and is not designed to be a neutral midpoint between text and image. To test whether either configuration is inherently easier, we flip the setup: swapping text and image components. Results (Table 12) show 1–5% differences, with no consistently easier design across MLLMs.

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 [14]			AI2-ARC [9]			GSM-Sym [25]			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 [14]			AI2-ARC [9]			GSM-Sym [25]			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

Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	82.9	81.4	83.0	83.0	0.0	90.0
InternVL3 (2B)	61.6	59.0	56.6	55.4	0.0	79.6
Mistral-Small	84.1	78.9	82.5	82.6	0.0	91.3
Phi-3.5	61.1	41.5	32.5	39.3	0.0	75.6
Phi-4	47.0	44.2	31.4	33.6	0.0	80.1
Qwen-2.5 (32B)	83.3	84.1	83.5	80.0	0.0	90.0
Qwen-2.5 (7B)	72.6	71.8	72.9	72.9	0.0	82.4
Haiku-4.5 (Claude)	90.1	87.3	85.1	85.2	0.0	93.7
Deepseek-Small	54.9	54.9	51.2	38.1	0.0	75.2
Deepseek-Tiny	29.5	26.6	25.3	17.6	0.0	68.7
Gemini-2.5-flash-lite	81.6	79.7	78.4	78.2	0.0	88.1
Gemma-3 (12B)	77.6	76.7	72.2	72.2	0.0	87.1
Gemma-3 (4B)	68.0	64.7	57.2	57.5	0.0	81.3
GPT-4o-mini	84.4	77.2	77.0	71.9	0.0	89.8
GPT-5-mini	89.3	86.5	87.8	89.0	0.0	93.4

GSM8K-Symbolic

Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	91.2	91.9	93.0	92.5	0.0	96.8
InternVL3 (2B)	52.1	40.3	53.1	64.6	0.0	76.8
Mistral-Small	92.6	91.7	93.3	93.2	0.0	96.5
Phi-3.5	66.9	42.9	48.3	45.2	0.0	78.3
Phi-4	76.5	55.8	54.5	60.7	0.0	85.9
Qwen-2.5 (32B)	93.6	93.1	93.0	93.4	0.0	95.7
Qwen-2.5 (7B)	84.4	82.0	82.1	83.0	0.0	91.9
Haiku-4.5 (Claude)	95.6	94.9	94.1	92.8	0.0	97.7
Deepseek-Small	67.3	54.7	60.9	57.5	0.0	80.4
Deepseek-Tiny	15.5	0.8	0.7	0.3	0.0	16.0
Gemini-2.5-flash-lite	79.4	60.1	50.5	59.5	0.0	84.8
Gemma-3 (12B)	91.3	87.7	88.1	89.5	0.0	95.1
Gemma-3 (4B)	82.2	70.5	67.4	70.0	0.0	89.6
GPT-4o-mini	91.4	86.7	89.0	88.6	0.0	95.2
GPT-5-mini	95.1	93.7	94.1	95.0	0.0	97.4

SOEBENCH

Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	73.3	70.0	72.7	71.3	0.0	88.7
InternVL3 (2B)	14.0	14.7	33.3	9.3	0.0	42.0
Mistral-Small	60.7	62.0	62.0	63.3	0.0	82.0
Phi-3.5	1.3	1.3	0.7	0.7	0.0	3.3
Phi-4	13.3	13.3	17.3	13.3	0.0	29.3
Qwen-2.5 (32B)	87.2	87.9	75.2	85.2	0.0	96.0
Qwen-2.5 (7B)	48.0	46.0	48.7	36.7	0.0	71.3
Haiku-4.5 (Claude)	96.7	95.3	97.3	96.0	0.0	99.3
Deepseek-Small	8.0	7.3	4.7	3.3	0.0	16.0
Deepseek-Tiny	0.7	0.0	0.0	0.0	70.5	0.7
Gemini-2.5-flash-lite	42.7	34.7	49.3	44.0	0.0	73.3
Gemma-3 (12B)	84.0	72.7	76.0	70.7	0.0	91.3
Gemma-3 (4B)	75.2	58.9	57.4	55.8	0.0	84.5
GPT-4o-mini	65.3	61.3	62.0	64.7	0.0	83.3
GPT-5-mini	98.7	95.3	97.3	98.0	0.0	100.0

AI2-ARC

Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	92.9	92.9	94.2	94.2	0.0	95.7
InternVL3 (2B)	76.0	74.0	74.0	72.8	0.0	89.7
Mistral-Small	92.7	90.0	92.6	92.0	0.0	95.8
Phi-3.5	73.3	54.3	37.9	57.0	0.0	83.6
Phi-4	59.9	48.5	35.0	44.1	0.0	86.1
Qwen-2.5 (32B)	92.9	92.9	92.9	90.4	0.0	95.2
Qwen-2.5 (7B)	86.1	87.0	88.3	88.9	0.0	93.1
Haiku-4.5 (Claude)	93.3	93.3	92.7	92.0	0.0	95.7
Deepseek-Small	69.1	68.7	70.0	57.2	0.0	85.3
Deepseek-Tiny	34.2	28.7	27.5	19.7	0.0	70.6
Gemini-2.5-flash-lite	90.4	89.0	88.8	89.0	0.0	93.2
Gemma-3 (12B)	90.5	90.2	88.6	89.6	0.0	94.4
Gemma-3 (4B)	79.2	78.2	71.4	72.7	0.0	90.5
GPT-4o-mini	93.6	89.3	89.9	87.8	0.0	95.4
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

Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	82.4	81.0	82.5	82.6	0.5	89.7
InternVL3 (2B)	61.2	58.5	56.0	55.0	1.1	79.4
Mistral-Small	83.9	78.7	82.3	82.3	0.0	91.3
Phi-3.5	58.1	40.3	31.3	37.1	4.5	72.8
Phi-4	45.0	43.2	30.6	32.5	0.6	79.1
Qwen-2.5 (32B)	83.3	84.1	83.3	79.7	0.1	90.0
Qwen-2.5 (7B)	72.4	71.5	72.7	72.7	0.0	82.3
Haiku-4.5 (Claude)	90.0	87.2	85.1	85.2	0.1	93.7
Deepseek-Small	54.8	54.6	51.1	37.7	0.2	75.1
Deepseek-Tiny	29.5	26.2	25.2	17.5	1.3	68.6
Gemini-2.5-flash-lite	81.3	79.4	78.0	77.6	0.1	87.9
Gemma-3 (12B)	77.4	76.5	71.9	71.8	0.2	87.0
Gemma-3 (4B)	64.0	60.2	53.5	53.8	3.6	76.8
GPT-4o-mini	84.2	77.0	76.7	71.7	0.0	89.6
GPT-5-mini	89.2	86.5	87.7	89.0	0.0	93.4

GSM8K-Symbolic

Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	90.6	91.4	92.6	92.2	0.1	96.6
InternVL3 (2B)	50.8	38.7	51.8	62.2	3.8	74.8
Mistral-Small	92.2	91.3	92.9	92.8	0.0	96.3
Phi-3.5	65.5	41.9	46.8	43.7	0.7	76.9
Phi-4	75.5	55.1	53.7	60.1	0.1	85.1
Qwen-2.5 (32B)	93.2	92.8	92.7	92.5	0.1	95.4
Qwen-2.5 (7B)	83.6	81.2	81.3	82.1	0.0	91.1
Haiku-4.5 (Claude)	95.4	94.4	93.6	92.4	0.1	97.7
Deepseek-Small	67.2	54.6	60.7	57.2	0.1	80.2
Deepseek-Tiny	15.2	0.8	0.7	0.3	0.4	15.8
Gemini-2.5-flash-lite	78.8	59.6	50.1	59.2	0.0	84.2
Gemma-3 (12B)	90.9	87.2	87.7	89.0	0.3	95.0
Gemma-3 (4B)	81.1	68.1	65.4	68.2	1.8	88.4
GPT-4o-mini	91.5	86.6	88.8	88.3	0.0	95.3
GPT-5-mini	94.9	93.5	93.8	94.5	0.0	97.3

SOEBENCH

Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	73.3	70.0	72.7	71.3	0.0	88.7
InternVL3 (2B)	14.0	14.7	33.3	9.3	0.0	42.0
Mistral-Small	60.7	62.0	62.0	63.3	0.0	82.0
Phi-3.5	1.3	1.3	0.7	0.7	0.0	3.3
Phi-4	13.3	13.3	17.3	13.3	0.0	29.3
Qwen-2.5 (32B)	86.7	87.3	74.7	84.7	0.0	95.3
Qwen-2.5 (7B)	48.0	46.0	48.7	36.7	0.0	71.3
Haiku-4.5 (Claude)	96.7	95.3	97.3	96.0	0.0	99.3
Deepseek-Small	8.0	7.3	4.7	3.3	0.0	16.0
Deepseek-Tiny	0.7	0.0	0.0	0.0	70.5	0.7
Gemini-2.5-flash-lite	42.7	34.7	49.3	44.0	0.0	73.3
Gemma-3 (12B)	84.0	72.7	76.0	70.7	0.0	91.3
Gemma-3 (4B)	74.0	55.3	52.7	52.0	1.2	82.7
GPT-4o-mini	65.3	61.3	62.0	64.7	0.0	83.3
GPT-5-mini	98.7	95.3	97.3	98.0	0.2	100.0

AI2-ARC

Model	Text	Mixed	Image	OCR First	CER	MMC
InternVL3 (14B)	93.0	93.2	94.3	94.3	0.2	95.8
InternVL3 (2B)	75.8	73.6	73.5	72.3	1.7	89.5
Mistral-Small	92.7	90.0	92.6	92.0	0.0	95.8
Phi-3.5	74.0	54.7	37.6	55.5	5.3	84.2
Phi-4	59.9	48.3	35.0	44.0	0.1	86.0
Qwen-2.5 (32B)	93.0	93.0	92.9	90.4	0.0	95.2
Qwen-2.5 (7B)	86.1	87.1	88.3	88.9	0.0	93.1
Haiku-4.5 (Claude)	93.3	93.3	92.7	92.0	0.0	95.7
Deepseek-Small	69.1	68.8	70.0	57.2	0.0	85.4
Deepseek-Tiny	34.6	28.9	27.1	19.6	0.5	70.6
Gemini-2.5-flash-lite	90.4	89.0	88.9	89.0	0.2	93.2
Gemma-3 (12B)	90.6	90.2	88.6	89.6	0.0	94.4
Gemma-3 (4B)	77.9	76.7	70.4	71.2	1.4	89.6
GPT-4o-mini	93.6	89.3	90.0	87.7	0.0	95.4
GPT-5-mini	94.6	92.7	93.3	94.4	0.0	96.0

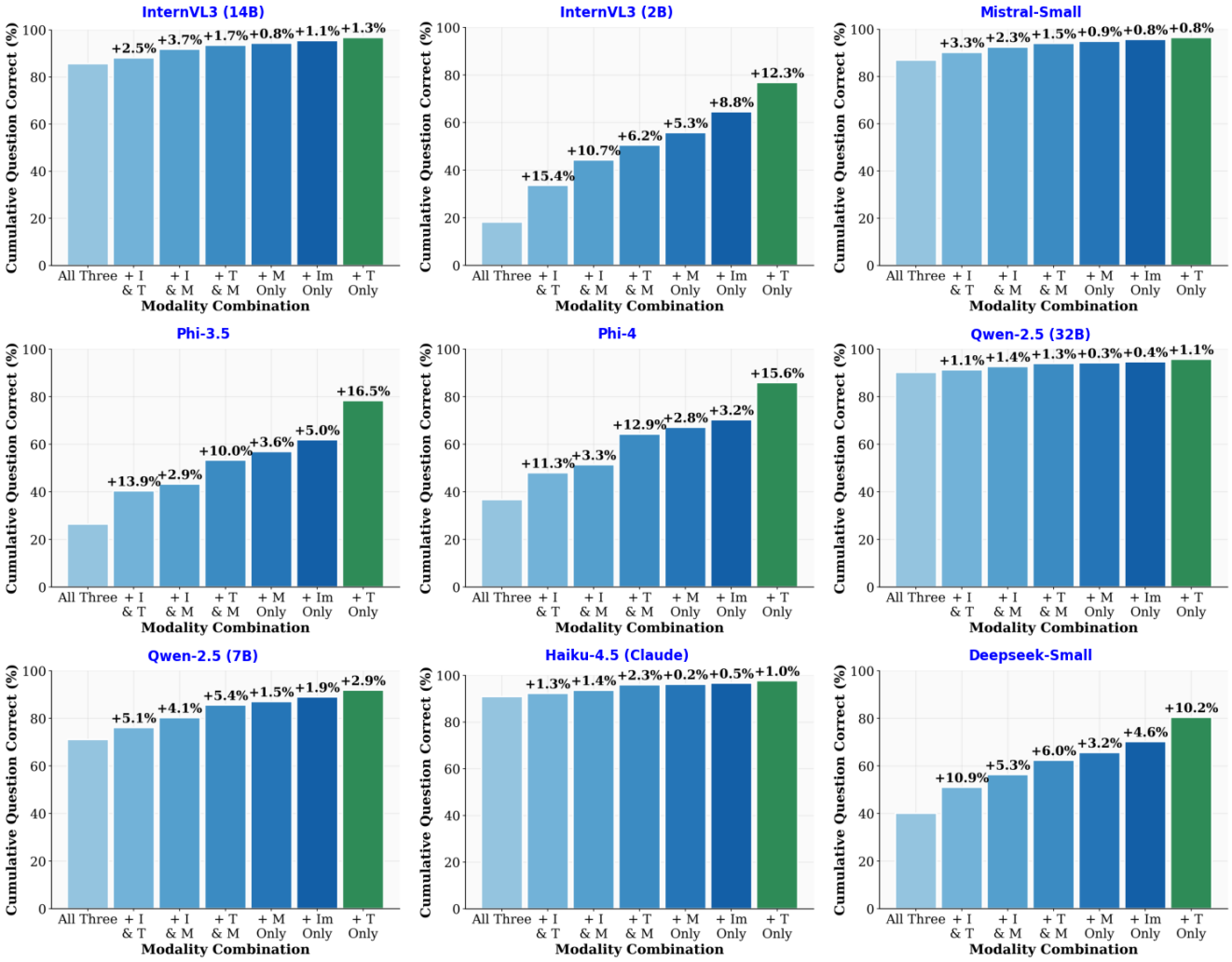


Figure 12. Cumulative distribution of correctly solved questions across modality combinations for models 1-8. Each step represents questions solvable in progressively fewer modalities, with the leftmost portion showing questions solved consistently across all three modalities.

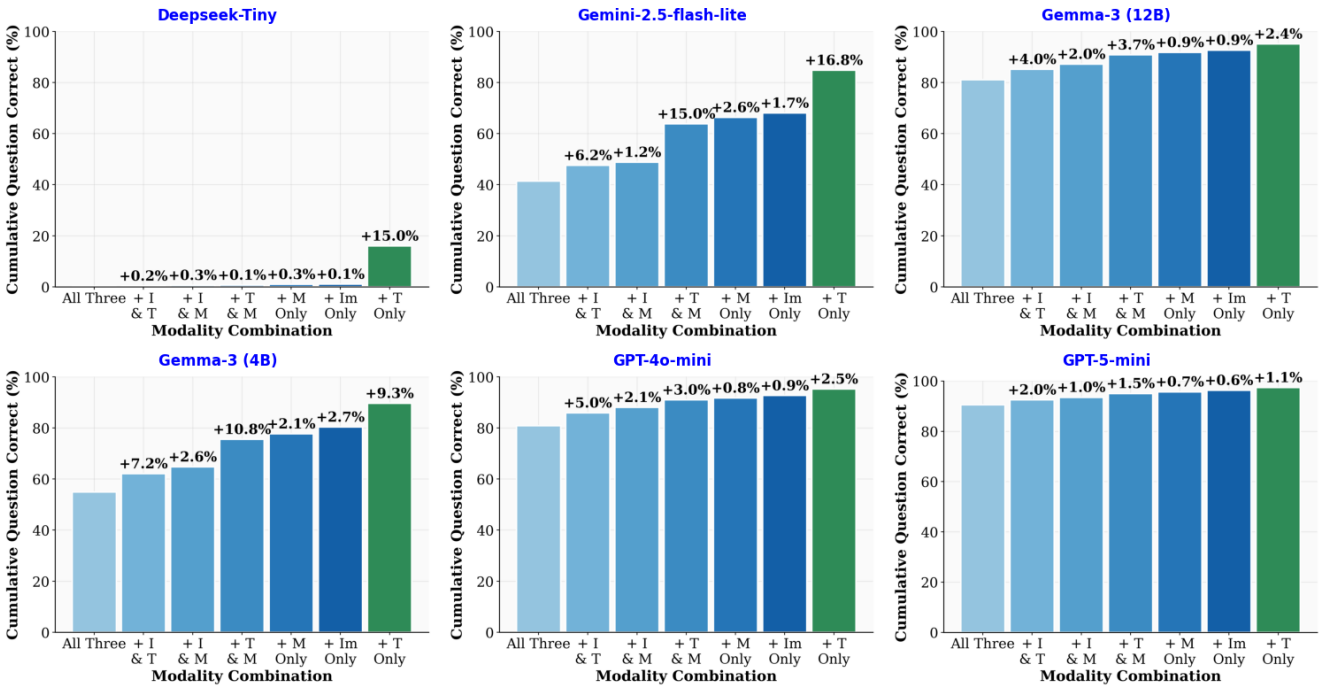


Figure 13. Cumulative distribution of correctly solved questions across modality combinations for models 9-15. Models with higher cross-modal consistency show larger proportions of questions solved in all modalities (leftmost region).

Table 11. GPT-5-mini shows the least inconsistency.

Model	**REST** (OCR✓)		**REST** (All Questions)		
	RER ↑	CFR ↓	RER ↑	CFR ↓	OCR✓↑
InternVL3 (14B)	78.4	19.6	78.1	19.9	95.3
InternVL3 (78B)	80.8	16.8	80.7	17.0	92.5
Qwen-2.5 (72B)	83.7	13.5	83.7	13.6	**99.4**
GPT-5-mini	**90.7**	**8.7**	**90.7**	**8.8**	99.0

Table 12. Flipping modality changes scores slightly (OCR correct).

Model	**AI2-ARC**		**GSM-Sym**		**MMLU**	
Accuracy mixed ↑	Orig	Flip	Orig	Flip	Orig	Flip
Qwen2.5 (7B)	**87.0**	81.3	**82.0**	78.9	**71.8**	67.3
Gemma (12B)	**90.2**	86.0	87.7	**87.9**	**76.7**	74.1
InternVL3 (14B)	**92.9**	91.3	**91.9**	89.2	**81.4**	76.8
GPT-5-mini	92.7	**93.9**	93.7	**94.2**	86.5	**88.6**

10.2. REST+

Cross-modal Consistency Tables 13 and 14 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 13. **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 × 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 15 and 16 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 17 and 18 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 14. **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 15. **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		
DPI	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 19 and 20 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 colored 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 21. As mentioned in the paper, we see that fewer text tokens are needed to achieve the same level of accuracy, or more

Table 16. **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		
DPI	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 17. **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	Curs.
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 [42], see Figure 14. The MMMU scores were found on public benchmarks[2]. Generally, we do find such a correlation, both for REST and REST+. Interestingly, the correlation seems low for models with high MMMU scores on REST+.

[2] https://mmmu-benchmark.github.io/#leaderboard, https://www.anthropic.com/news/claude-haiku-4-5, 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.

Table 18. **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	Curs.
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 19. **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	C
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

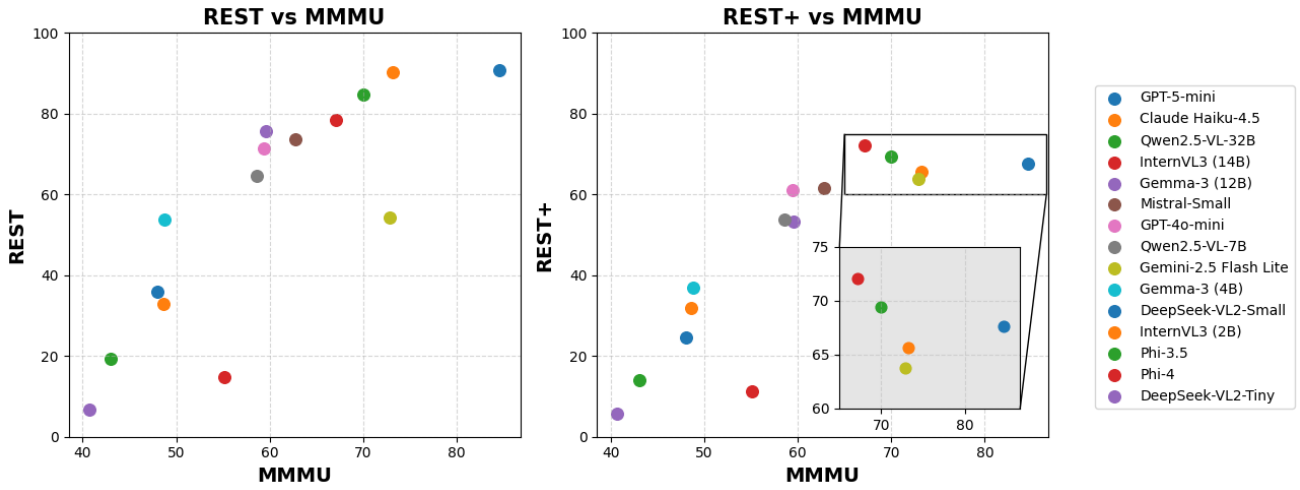


Figure 14. Models that perform well on MMMU also score well on REST and REST+. The zoom inset shows that models with high MMMU scores do not generally obtain high REST+ scores.

Table 20. **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

10.3. Natural chess images instead of rendered text

It is important to examine whether inconsistency extends beyond typographic inputs. We construct an additional same-content setting using chess positions, where spatial information translates naturally to text. Using ChessReD (Masouris 2023, DOI), we ask yes-or-no questions about positions presented as (a) natural image, (b) generated image, or (c) text. Results (Table 22) confirm that cross-modal inconsistency extends to natural images: answers vary by

Table 21. **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	DPI 200
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

(a) Natural image (3D)

(b) Generated image (2D)

```
'white:  e5 pawn, e4
pawn, f4 pawn, a3
knight, g3 pawn, d2
rook, h2 pawn, a1
rook, f1 bishop, g1
king ; black:  b8
rook, c8 rook, g8
king, f7 pawn, g7
pawn, a6 pawn, b4
knight, g4 bishop, h3
pawn'
```

(c) Text

Figure 15. Q1: Is a white pawn on the same row as another white pawn? (Y). Q2: Can the white bishop capture a black pawn in a single move? (Y).

modality even when semantic understanding is verified (by querying the board state). Across modalities, the accuracy can differ max. 35% (GPT-5-mini). Across models, RER ranges from 58.5 to 77.5, and models performing well on REST (GPT-5-mini, Mistral, Qwen-32B) also perform well on chess.

Table 22. Cross-modal inconsistency for chess. Sorted by RER.

Model	**Chess**					**REST**	
	3D ↑	2D ↑	Text ↑	RER ↑	CFR ↓	RER ↑	CFR ↓
InternVL3 (14B)	58.4	60.1	75.8	58.5	46.8	78.4	19.6
Gemma-3 (4B)	57.3	49.9	66.8	59.0	49.1	53.9	42.3
Gemma-3 (12B)	60.7	61.7	72.5	65.4	40.9	75.8	21.3
Qwen2.5 (32B)	**64.9**	69.4	85.3	65.5	35.7	84.7	13.6
Mistral-Small	62.6	74.2	75.7	67.7	36.7	73.6	23.9
GPT-5-mini	61.5	**94.0**	**96.4**	**77.5**	**22.9**	**90.7**	**8.7**