Acknowledgements

We thank Ross Girshick for discussing the idea behind this project. We also thank OpenAI's Researcher Access Program (ID 0000005368) for granting us API access. Serena Yeung-Levy is a Chan Zuckerberg Biohub — San Francisco Investigator.

Broader Impacts and Ethics Statement

AutoConverter and *VMCBench* simplify and standardize vision language model (VLM) evaluation, providing a valuable tool for advancing VLM development through scalable, consistent, and reproducible evaluation. Beyond VLM evaluation, *AutoConverter* can be applied to other domains, such as education, to generate challenging and high-quality questions. However, human verification is essential to ensure the generated questions align with their intended purpose, maintain proper value, and achieve the appropriate level of difficulty without introducing biases.

Reproducibility Statement

We provide an open-source implementation of our work at https://github.com/yuhui-zh15/AutoConverter. This will enable researchers to reproduce all the experiments in paper and conduct their own analyses.

Limitations

AutoConverter achieves high question correctness, but 5% of the questions with the highest correctness scores still contain errors, although half of which are due to flaws in the original datasets. Moreover, while *VMCBench* includes a diverse range of models and datasets, its coverage remains incomplete. Moving forward, we aim to leverage *AutoConverter* to expand dataset and model coverage, addressing evolving needs in VLM evaluation.

Summary of Appendix

- In §A, we present further analysis of the evaluation failures for open-ended questions discussed in §3.
- In §B, we provide supplementary details on the *AutoConverter* framework described in §4.
- In §C, we elaborate more details of *VMCBench* as outlined in §5.

A. Supplementary Section 3

A.1. Examples of Rule-based Evaluation Failures

In the main paper, we demonstrated that rule-based evaluation of open-ended questions leads to poor evaluation outcomes. Table 3 provides six examples of rule-based evaluation failures, clearly illustrating that these methods fail to account for semantic similarity and penalize formatting errors, resulting in highly inaccurate evaluation results.

A.2. Examples of Model-based Evaluation Failures

In the main paper, we demonstrated that model-based evaluation is sensitive to model versions. To investigate this further, Table 4 presents six examples of inconsistencies caused by different model-based evaluations. We observe that GPT-4o-0806 often assigns a perfect score of 1.0 for similar predictions and answers, whereas GPT-4o-0513 tends to assign a score of 0.9. This behavior variation introduces significant differences in evaluation results and raises concerns about reproducibility in future research.

B. Supplementary Section 4

B.1. Prompts for *AutoConverter*

In the main paper, we introduced the agentic system of *AutoConverter*, comprising the proposer, reviewer, selector, evaluator, and refiner agents. This system creates a large pool of high-quality distractor options to increase question difficulty while ensuring correctness in the converted multiple-choice questions.

Detailed prompts for the proposers, designed to create distractors addressing *vision*, *reasoning*, *data*, *concept*, and *bias* errors, are shown in Figures 11, 12, 13, 14, and 15, respectively. These prompts are carefully crafted to align with common types of human errors, as defined in the corresponding sections.

Figure 16 presents prompts for the *reviewer*, whose feedback iteratively refines the distractors to improve their quality. The *selector* prompts in Figure 17 guide the selection of the three most challenging distractors to enhance question difficulty.

Figures 18 and 19 show prompts for the *evaluator* and *refiner*, respectively. This iterative process ensures the correctness of the generated questions, guaranteeing that there is only one correct answer.

B.2. *AutoConverter* Results for Additional Datasets

In the main paper, we demonstrated *AutoConverter*'s ability to generate challenging multiple-choice questions for three datasets: MMMU, MathVista, and AI2D. Figure 7 extends this analysis to five additional datasets with human-created distractors: A-OKVQA, RealWorldQA, ScienceQA, SEEDBench, and MMStar. Across these datasets, VLMs consistently achieved similar or lower accuracy on *AutoConverter*-generated questions compared to the original ones, demonstrating the system's capability to produce highly challenging multiple-choice questions.

B.3. *AutoConverter* Results for Different Models

In the main paper, we constructed *AutoConverter* using GPT-4o. A potential concern is whether this choice introduces bias into the generated questions.

To examine this, we used three state-of-the-art proprietary VLMs—GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro—to generate questions. We evaluated various VLMs on these questions and computed the correlations of their performance rankings. If no bias exists, we expect high correlations across all question sets, as they should reflect the true discriminative power of the models.

Figure 8 supports our hypothesis. The rankings derived from GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro questions exhibit near-perfect correlations (0.90 Spearman correlation averaged over three datasets), indicating that the choice of generator does not introduce significant bias.

B.4. Correlation between Open-ended and Multiple-choice Questions

In the main paper, we highlighted the challenges of accurately evaluating open-ended questions. Therefore, we propose an alternative solution to convert open-ended questions into a multiple-choice format. A key question here is whether converting open-ended questions into multiple-choice format preserves their discriminative power.

To address this, we treat model-based evaluation of open-ended questions as a proxy for ground-truth evaluation. Specifically, if a model-based evaluator determines that model A outperforms model B, we consider this ranking correct. This assumption is valid because model-based evaluations have a high correlation with human judgments, albeit with instability across versions. By using the same GPT version as the evaluator, we eliminate such instability.

We compare the correlation between model-based evaluation of open-ended questions and rule-based evaluation of multiple-choice questions against the correlation between model-based and rule-based evaluation of open-ended questions. Our findings, shown in Figure 9, reveal that the correlation for multiple-choice questions is significantly higher—0.85, 0.71, and 0.97 for VQAv2, OKVQA, and VizWiz, respectively—compared to rule-based open-ended evaluations, which achieve correlations of 0.09, 0.19, and 0.00. This demonstrates that converting to multiple-choice questions improves evaluation accuracy and retains discriminative power.

C. Supplementary Section 5

C.1. Human Evaluation Results on *VMCBench*

VMCBench provides a scalable, consistent, and reproducible benchmark for evaluating and advancing VLMs. The current best-performing model, GPT-4o, achieves an accuracy of 80.6%.

To assess the remaining room for improvement, we conducted a human evaluation of *VMCBench*. Human experts achieved an accuracy of 91.7%, highlighting significant opportunities for further model improvement. Among the 8.3% human errors, approximately three-quarters of the questions require extensive knowledge, while a quarter are ambiguous and unanswerable.

C.2. Full Evaluation Results on *VMCBench*

In the main paper, we reported model performances grouped by benchmarked capabilities. Table 5 presents the full evaluation results for 33 VLMs across 20 individual benchmarks, further validating our conclusions.

C.3. Scaling Trends on *VMCBench*

Scaling laws are a cornerstone of model development, demonstrating that larger models typically achieve better performance. To explore this, we plotted the scaling trends of VLM families in log scale based on known model sizes, as shown in Figure 10.

Surprisingly, we observe a clear **log-linear** scaling trend across most VLM families, indicating that *VMCBench* offers a smooth evaluation gradient for varying capabilities. Certain model families outperform others, leaving further exploration of these scaling trends to future work.

C.4. Option Permutation Sensitivity on *VMCBench*

To assess whether VLMs are sensitive to option permutation, we randomly shuffled all correct options twice using two different random seeds. We observed no significant difference in performance on *VMCBench* (Table 2). This result demonstrates the robustness of multiple-choice evaluation in *VMCBench*.

Question	M_1	M_2	M_3	M_4	M_5	M_6
Original	59.7	52.6	68.3	78.8	78.1	80.8
Shuffled 1	59.8	52.7	68.0	78.6	77.7	81.0
Shuffled 2	59.9	52.5	67.9	78.7	78.2	80.3

Table 2. **Performance on *VMCBench* with shuffled options.** M_1 is PaliGemma-3B, M_2 is LLaVA-1.5-7B, M_3 is Phi-3.5-Vision, M_4 is Qwen2-VL-7B, M_5 is Claude-3.5-Sonnet, M_6 is GPT-4o.

C.5. Additional Examples from *VMCBench*

To provide a deeper understanding of *VMCBench*, we include 60 examples (three from each of the 20 sources) in Tables 6 through 12. These examples demonstrate the high quality of our dataset and offer valuable insights for readers.

Image	Question	Answer	Prediction	VQAScore	ModelScore
	Where is the man's left hand holding?	chin	On his chin.	0	1
	How many sinks are there?	2	Two sinks.	0	1
	What is the light source?	street lights	Streetlights	0	1
	What is the horse jumping over?	log	Log fence	0	1
	Where is the train going?	svanemollen	Svanemøllen	0	1
	What colors are on this train?	blue and yellow and silver	Blue, yellow, and silver.	0	1

Table 3. **Examples of rule-based evaluation failures of open-ended questions.** Rule-based methods fail to account for semantic similarity and penalize formatting errors, resulting in highly inaccurate evaluation results.

Image	Question	Answer	Prediction	GPT-4o-0513 Score	GPT-4o-0806 Score
	Does the giraffe appear dangerous?	no	The giraffe in the image does not appear to be dangerous. It seems to be calm and at ease with the person's presence, indicating a level of trust and familiarity between them.	0.9	1.0
	Is the man happy?	yes	The man in the image appears to be happy. His facial expression, characterized by a broad smile and eyes that seem to be crinkling at the corners, typically indicates joy or amusement...	0.9	1.0
	Is that wine in the background?	no	No, there is no wine in the background. The image shows a pizza and a can of Coca-Cola on a desk, with a laptop and some other items in the background.	0.9	1.0
	What are all the scene text in the image?	S3302CD	The scene text in the image reads 'S 3302 CD'.	0.9	1.0
	What is this banknote?	ten pound or 10 pound or British pound	This is a ten-pound banknote from the United Kingdom, featuring the portrait of Queen Elizabeth II.	0.9	1.0
	How many birds are there?	6	There are six birds perched on the branch, each displaying a unique coloration and size, which suggests they may belong to different species.	0.9	1.0

Table 4. **Examples of model-based evaluation failures of open-ended questions.** GPT-4o-0806 often assigns a perfect score of for similar predictions and answers, whereas GPT-4o-0513 tends to assign a score of 0.9. This behavior variation introduces significant differences in evaluation results and raises concerns about reproducibility in future research.

Model	SEEDBench	MMStar	A-OKVQA	VizWiz	MMVet	VQAv2	OKVQA	MMMU	MathVista	ScienceQA	
Qwen2-VL-72B	84.2	71.5	93.4	94.9	88.5	92.4	94.8	70.2	70.3	87.1	
GPT-4o	84.2	62.2	92.0	94.4	80.6	88.7	94.1	70.0	51.0	86.9	
Molmo-72B	82.0	62.5	88.5	88.5	79.1	85.6	94.1	59.4	60.9	89.4	
Qwen2-VL-7B	82.7	60.3	90.1	92.4	82.0	91.4	92.6	54.1	48.0	86.9	
Claude-3.5-Sonnet	78.3	53.7	86.4	87.2	87.8	84.7	91.1	59.6	56.9	79.9	
Cambrian-34B	83.5	59.4	91.1	90.7	81.3	88.4	91.9	55.0	60.9	83.5	
Gemini-1.5-Pro	77.3	52.0	88.0	89.0	79.9	80.8	90.4	59.6	56.9	83.0	
VILA1.5-40B	81.2	58.0	90.8	90.2	77.0	87.3	93.3	58.2	65.3	83.3	
GPT-4o-Mini	79.3	47.7	86.1	92.2	80.6	88.2	92.1	56.5	43.1	78.7	
Qwen2-VL-2B	77.8	44.7	86.6	88.2	70.5	88.7	88.6	46.2	39.1	76.0	
CogVLM2-19B	77.3	48.2	87.8	87.3	73.4	85.2	87.4	39.7	35.6	90.5	
Phi-3-Vision	78.3	46.8	81.6	79.9	67.6	79.2	85.2	44.7	38.6	92.1	
Cambrian-13B	79.3	48.5	86.4	87.3	76.3	87.0	90.6	41.3	39.6	77.4	
Cambrian-8B	78.3	53.0	84.5	88.0	68.3	85.6	87.9	43.3	45.5	77.4	
Molmo-7B-D	74.1	46.8	82.4	81.9	63.3	80.3	83.5	43.0	37.6	91.9	
Idefics2-8B	77.8	49.4	85.4	84.8	69.8	86.8	90.4	38.5	41.6	91.9	
Molmo-7B-O	75.1	45.8	80.5	78.9	66.2	78.2	83.2	44.0	35.6	90.0	
Phi-3.5-Vision	74.8	45.8	76.5	75.7	64.0	79.4	83.5	45.2	39.1	87.3	
VILA1.5-13B	77.5	44.9	81.9	83.3	63.3	82.4	88.6	42.8	48.5	72.6	
DeepSeek-VL-7B	75.6	39.9	80.5	82.1	63.3	83.6	87.7	41.1	33.2	86.7	
Molmo-1B	70.9	43.0	77.9	80.9	54.7	75.7	83.0	36.3	27.7	89.1	
CogVLM-17B	70.9	42.8	80.7	85.0	59.7	80.1	86.7	37.5	36.1	66.7	
VILA1.5-8B	74.3	40.9	77.9	79.4	64.7	80.8	88.6	38.0	49.0	71.5	
Gemini-1.5-Flash	56.3	38.0	67.0	68.5	53.2	64.1	70.6	48.1	57.9	66.3	
PaliGemma-3B	74.6	39.9	87.3	77.0	50.4	87.7	85.2	29.1	30.7	94.3	
VILA1.5-3B	74.3	38.5	76.9	80.9	57.6	78.5	85.4	34.4	39.6	64.9	
DeepSeek-VL-1.3B	70.9	37.5	74.4	82.1	52.5	80.3	84.7	31.0	22.3	63.8	
LLaVA1.5-13B	66.2	37.3	76.5	76.0	59.7	64.1	85.2	37.5	31.7	66.3	
LLaVA1.5-7B	62.2	34.2	72.5	73.8	54.0	66.7	82.2	35.6	31.2	68.6	
Chameleon-30B	53.6	33.0	57.2	52.2	48.9	58.3	68.6	34.4	32.7	57.9	
InstructBLIP-7B	52.8	34.7	61.9	65.4	39.6	59.7	71.9	31.0	22.8	46.8	
InstructBLIP-13B	48.4	29.0	63.3	64.2	43.2	64.1	71.6	25.7	19.8	50.0	
Chameleon-7B	44.9	31.6	46.4	40.4	29.5	41.4	53.1	32.7	22.3	53.6	

Model	RealWorldQA	GQA	MathVision	TextVQA	OCRVQA	AI2D	ChartQA	DocVQA	InfoVQA	TableVQABench	Avg.
Qwen2-VL-72B	78.7	91.4	38.0	98.4	95.1	85.2	90.1	99.1	92.6	83.3	85.0
GPT-4o	75.0	84.4	34.4	97.5	95.3	81.8	80.0	98.2	79.0	76.4	80.3
Molmo-72B	70.0	83.1	36.9	95.3	94.0	79.7	81.2	94.7	74.9	75.2	78.7
Qwen2-VL-7B	67.9	89.0	30.6	97.3	95.6	75.2	78.9	98.0	78.6	69.8	78.1
Claude-3.5-Sonnet	63.2	81.9	35.1	93.6	93.2	72.9	87.9	97.7	79.1	85.6	77.8
Cambrian-34B	65.6	87.5	42.7	94.8	96.6	74.9	78.9	89.3	64.5	59.0	77.0
Gemini-1.5-Pro	60.6	79.0	49.2	91.5	93.8	72.2	72.9	83.1	64.3	70.7	74.7
VILA1.5-40B	63.1	88.8	33.3	91.2	95.1	71.3	73.9	77.1	57.1	57.9	74.7
GPT-4o-Mini	65.8	80.2	28.3	95.1	92.5	72.7	71.6	95.3	70.5	64.0	74.0
Qwen2-VL-2B	57.6	83.9	32.4	94.8	91.5	68.8	70.9	95.5	65.9	61.5	71.5
CogVLM2-19B	56.9	81.4	29.4	96.2	88.3	68.3	75.7	92.7	62.4	63.7	71.4
Phi-3-Vision	54.6	79.0	29.2	92.1	89.1	69.5	78.0	93.3	61.8	66.4	70.3
Cambrian-13B	58.3	86.8	24.5	91.0	93.5	71.8	71.6	82.6	51.8	55.4	70.0
Cambrian-8B	61.5	85.3	25.2	92.4	89.6	71.3	70.9	83.1	45.4	56.5	69.6
Molmo-7B-D	58.7	74.1	27.9	93.0	90.4	66.3	74.8	88.9	65.7	64.9	69.5
Idefics2-8B	55.0	80.0	27.6	91.7	93.8	66.3	67.0	81.5	47.0	47.3	68.7
Molmo-7B-O	55.0	75.3	26.1	90.8	86.3	67.4	74.3	84.4	58.3	59.9	67.8
Phi-3.5-Vision	47.9	81.2	30.8	82.9	91.5	73.1	73.6	80.0	53.7	62.6	67.4
VILA1.5-13B	48.2	83.6	29.0	81.6	89.1	65.8	55.3	55.9	35.7	38.5	63.4
DeepSeek-VL-7B	48.4	82.6	23.6	83.6	88.1	59.5	60.8	61.9	38.9	43.5	63.2
Molmo-1B	45.2	76.0	26.3	87.0	87.8	60.1	62.4	79.1	51.4	48.2	63.1
CogVLM-17B	44.0	78.2	29.9	65.4	90.2	59.5	61.5	65.7	40.6	45.7	61.3
VILA1.5-8B	45.0	79.2	27.9	75.5	88.1	61.3	46.3	49.7	34.3	40.8	60.7
Gemini-1.5-Flash	48.0	62.1	40.2	79.6	80.2	53.1	55.2	69.9	51.3	51.8	59.1
PaliGemma-3B	46.6	86.1	20.9	34.6	71.5	64.7	51.8	62.6	42.9	42.8	59.0
VILA1.5-3B	43.1	79.5	26.5	76.0	81.9	53.8	44.5	51.2	30.4	31.5	57.5
DeepSeek-VL-1.3B	42.4	76.5	25.6	78.9	80.1	47.6	47.7	53.9	33.6	36.3	56.1
LLaVA1.5-13B	34.6	81.2	25.8	67.4	84.2	56.0	29.8	38.5	30.4	30.0	53.9
LLaVA1.5-7B	35.1	72.6	25.4	64.3	83.7	47.4	28.0	37.2	34.3	27.9	51.8
Chameleon-30B	35.3	59.2	27.6	39.3	56.7	47.2	28.4	32.1	32.0	29.7	44.2
InstructBLIP-7B	30.7	59.4	20.2	50.8	44.6	38.3	24.5	31.0	32.0	23.9	42.1
InstructBLIP-13B	26.1	60.6	25.8	48.1	49.2	33.0	20.0	28.3	24.4	26.6	41.1
Chameleon-7B	29.1	44.5	26.1	35.7	49.0	36.0	24.1	26.7	32.0	29.1	36.4

Table 5. **Performance of 33 vision language models on 20 subsets of *VMCBench* test set.**

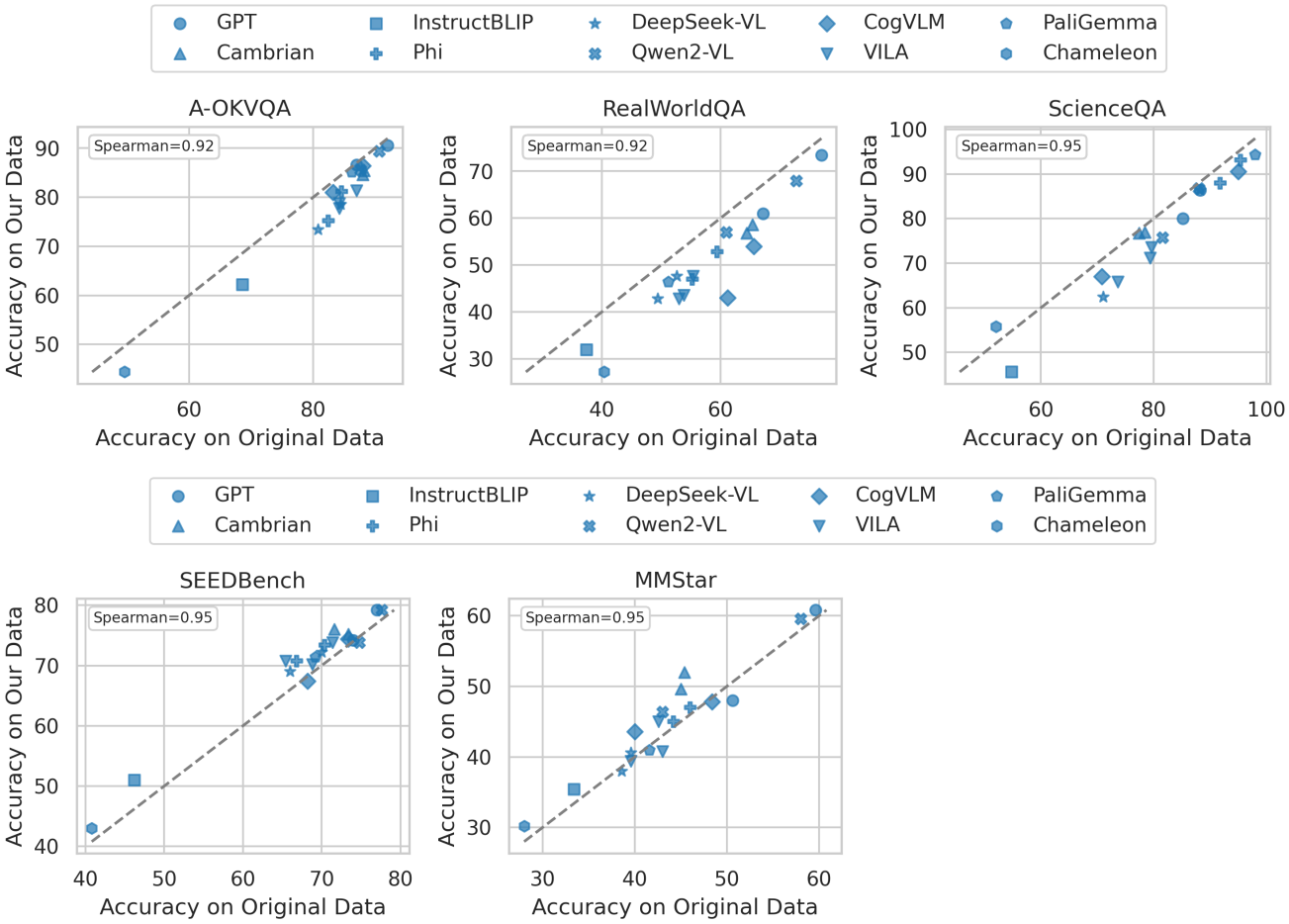


Figure 7. ***AutoConverter* generates challenging multiple-choice questions.** Using *AutoConverter*, we generated distractors for questions and answers from five multiple-choice datasets: A-OKVQA, RealWorldQA, ScienceQA, SEEDBench, and MMStar, and compared them with original human-created distractors. We evaluated various VLMs on both the *AutoConverter*-generated and the original questions, finding that VLMs consistently achieved similar or even lower accuracy on the *AutoConverter*-generated questions compared to the original ones.

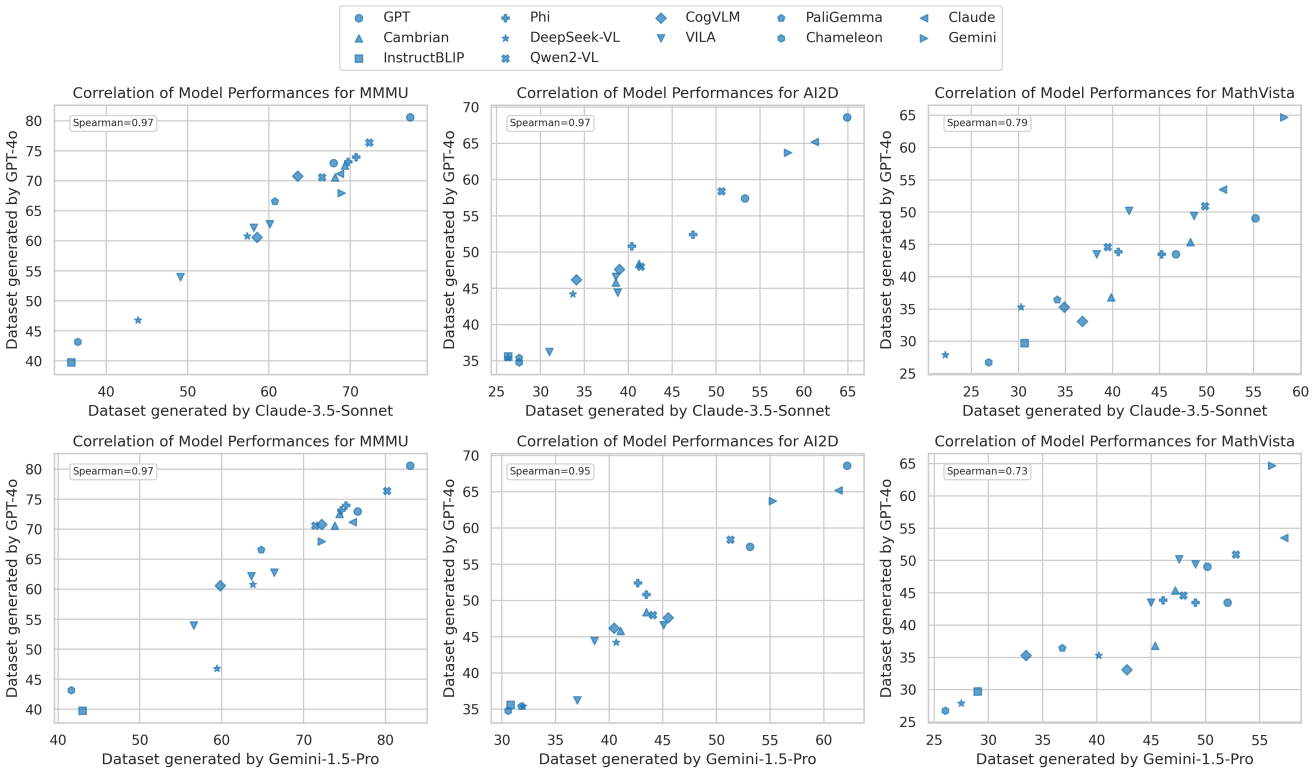


Figure 8. ***AutoConverter* results for different models.** To examine whether GPT-4o used in *AutoConverter* introduces model bias, we used three state-of-the-art proprietary VLMs—GPT-4o, Claude-3.5-Sonnet, and Gemini-1.5-Pro—to generate questions. We evaluated 18 VLMs on these questions and computed the correlations of their performance rankings. We observe high correlations across all question sets, indicating no model bias exists, and these questions reflect the true discriminative power of the models.

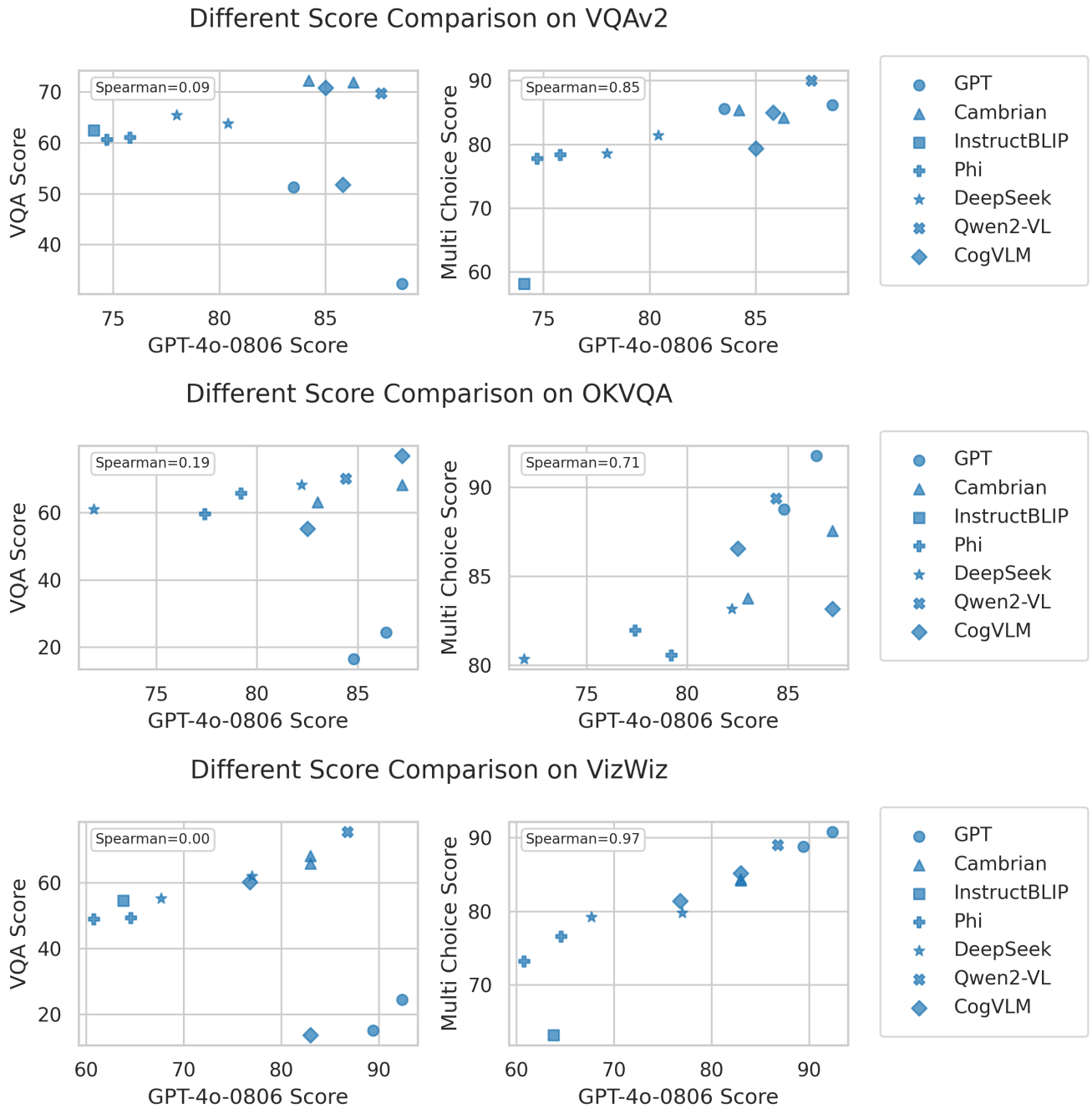


Figure 9. **Converting to multiple-choice questions improves evaluation accuracy and retains discriminative power.** We treat model-based evaluation of open-ended questions as a proxy for ground-truth evaluation. We compare the correlation between model-based evaluation of open-ended questions and rule-based evaluation of multiple-choice questions against the correlation between model-based and rule-based evaluation of open-ended questions. We find that the correlation for multiple-choice questions is significantly higher compared to rule-based open-ended evaluations, demonstrating that converting open-ended questions into multiple-choice format preserves their discriminative power and simplifies the evaluation.

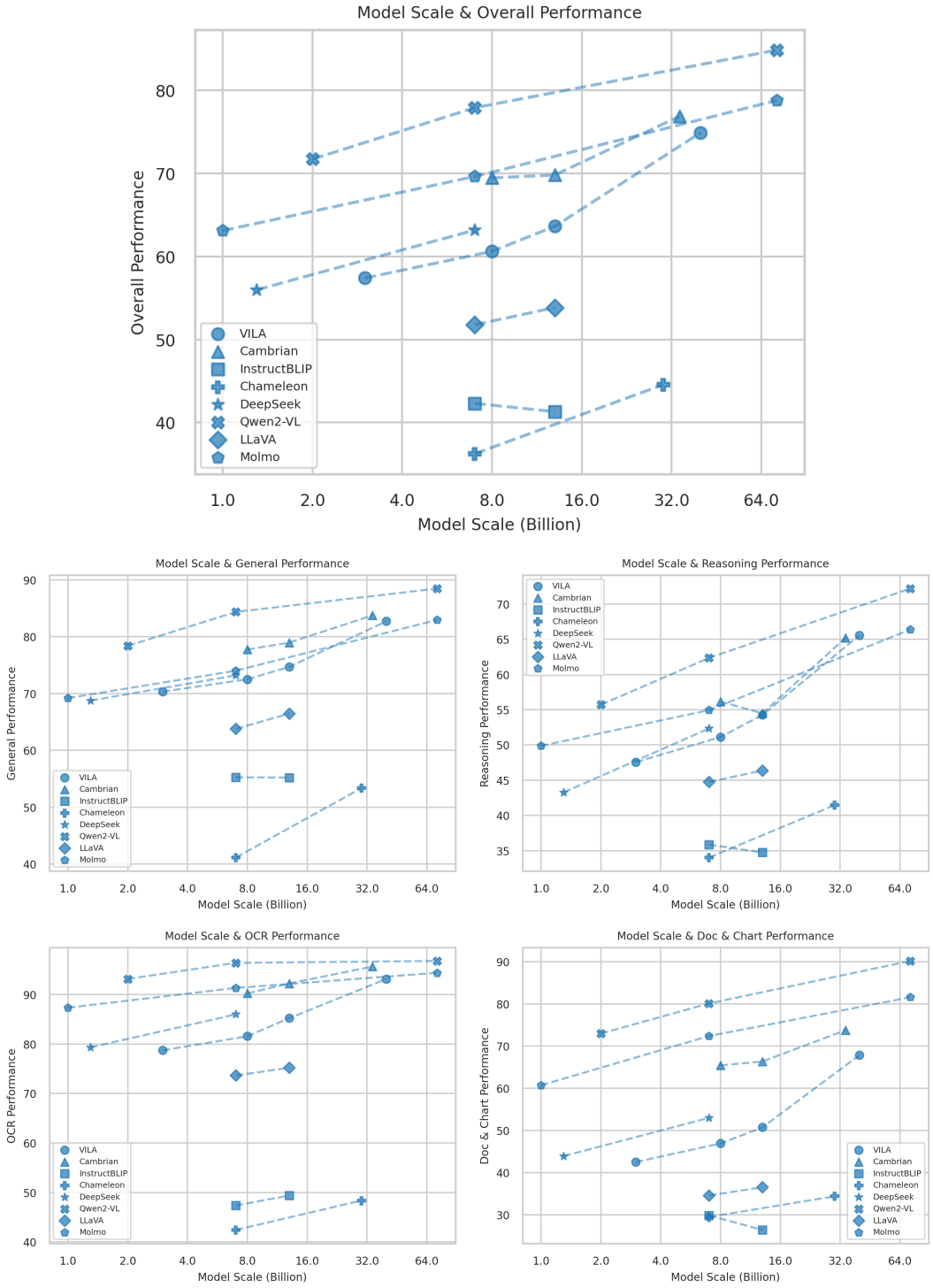


Figure 10. **Scaling trends on *VMCBench*.** We observe a clear **log-linear** scaling trend across most VLM families, indicating that *VMCBench* offers a smooth evaluation gradient for varying capabilities.

```
You are an expert in creating challenging and educational multiple-choice questions, specializing in visual
    ↪ interpretation errors.
Your task is to generate plausible but incorrect options (distractors) for given image-based question(s), focusing on
    ↪ misinterpretations of visual information.

Given:
1. One or more images
2. An open-ended question about the image(s)
3. The correct answer to the question

Your task:
1. Carefully analyze and understand the provided image(s). Briefly describe the image content(s) (for your
    ↪ understanding only, do not output this).
2. Generate {num_choice} unique and plausible distractor options based on visual interpretation errors. Each
    ↪ distractor should:
    - Be directly related to misinterpretation of the image(s)
    - Seem potentially correct at first glance
    - Be very misleading for students due to visual misunderstanding
    - Contain a subtle error in interpreting visual information that makes it incorrect
    - Vary in difficulty and the type of visual misinterpretation it represents

3. Ensure you understand how the correct answer relates to specific visual elements in the image(s).
4. Focus on common visual interpretation errors, including:
    - Misreading Graphs or Charts: Create options that misinterpret trends, scales, or relationships in visual data
    - Spatial Misinterpretation: Develop options that misunderstand spatial relationships or perspectives in the
        ↪ image(s)
    - Color Confusion: Include options that misinterpret color-coded information or subtle color differences
    - Pattern Misrecognition: Generate options that incorrectly identify or extend patterns in the image(s)
    - Detail Oversight: Create options that miss crucial details or focus on irrelevant visual elements
    - Scale Misjudgment: Include options that misinterpret the scale or proportions of elements in the image(s)
    - Cross-Image Miscomparison: When multiple images are provided, create options that incorrectly compare or
        ↪ contrast elements across images

5. Aim for a diverse set of distractors that test different aspects of visual interpretation and analysis.
6. Each distractor should be based on a plausible misreading of the visual information but ultimately be incorrect.
7. Consider the specific type(s) of image(s) (e.g., photograph, diagram, graph) and generate errors typical for those
    ↪ visual formats.
8. Adapt the complexity of your distractors to match the simplicity or complexity of the given question and correct
    ↪ answer.
9. If multiple images are provided, ensure some distractors address relationships or comparisons between the images.

10. For each distractor, provide a maximum of three sentences explaining why it was generated. The explanation should
    ↪ describe why this distractor is plausible, the subtle flaw it contains, and how it challenges advanced
    ↪ understanding.

Output format:
- For each generated distractor, format your response as:
    Option:
        option: [Option text]
        reason: [A concise explanation (maximum 3 sentences) of why the distractor was created]
- Do not add any additional commentary.

Remember:
- Your goal is to create challenging yet ultimately incorrect options that specifically target misinterpretations of
    ↪ visual information.
- All distractors should be plausible enough to be considered by a student who hasn't fully developed their visual
    ↪ literacy skills, but clear enough to be definitively incorrect upon careful visual analysis.
- Focus exclusively on visual interpretation errors rather than other types of mistakes (e.g., conceptual
    ↪ misunderstandings, reasoning errors).
- The distractors should directly relate to misunderstandings of the image(s) itself, not just the general topic of
    ↪ the question.
- Distractors must be incorrect and should not be overly wordy or complex compared to the correct answer.
- Ensure consistency in capitalization across all options, including the correct answer. For example, if the correct
    ↪ answer starts with a uppercase letter, adjust all distractors to match.
- When dealing with multiple images, consider how visual interpretation errors might arise from comparing or
    ↪ contrasting information across the images.
- Pay attention to any visual relationships, patterns, or differences that span multiple images, and create
    ↪ distractors that plausibly misinterpret these inter-image connections.
```

Figure 11. **Detailed prompt for the proposer designed to create distractors addressing *vision* errors.**

```
You are an expert in creating challenging and educational multiple-choice questions, specializing in reasoning errors.
Your task is to generate plausible but incorrect options (distractors) for given image-based question(s), focusing on
    ↪ flaws in logical reasoning and inference.

Given:
1. One or more images
2. An open-ended question about the image(s)
3. The correct answer to the question

Your task:
1. Carefully analyze and understand the provided image(s). Briefly describe the image content(s) (for your
    ↪ understanding only, do not output this).
2. Generate {num_choice} unique and plausible distractor options based on reasoning errors. Each distractor should:
    - Be related to the image(s) and question
    - Seem potentially correct at first glance
    - Be very misleading for students due to faulty reasoning
    - Contain a subtle logical flaw that makes it incorrect
    - Vary in difficulty and the type of reasoning error it represents

3. Ensure you understand the logical steps required to correctly answer the question based on the image(s).
4. Focus on common reasoning errors, including:
    - Complex Reasoning Flaws: Create options that require multi-step reasoning but contain logical gaps or invalid
        ↪ assumptions
    - Causal Inversion: Develop options that reverse cause and effect relationships
    - Context Neglect: Include options that ignore important contextual information provided in the question or
        ↪ image(s)
    - False Analogies: Generate options that draw incorrect parallels or comparisons
    - Hasty Generalizations: Create options that jump to conclusions based on insufficient evidence
    - Cross-Image Fallacies: When multiple images are provided, create options that make invalid logical connections
        ↪ or comparisons between images

5. Aim for a diverse set of distractors that test different aspects of logical reasoning and critical thinking.
6. Each distractor should follow a seemingly logical path but ultimately lead to an incorrect conclusion due to
    ↪ flawed reasoning.
7. If the question involves a specific subject area, consider common logical pitfalls or fallacies unique to that
    ↪ field.
8. If the question does not involve explicit reasoning, focus on creating plausible reasoning statements that could
    ↪ be mistakenly associated with the correct answer.
9. Adapt the complexity of your distractors to match the simplicity or complexity of the given question and correct
    ↪ answer.
10. If multiple images are provided, ensure some distractors address relationships or comparisons between the images,
    ↪ focusing on logical errors in interpreting these relationships.

11. For each distractor, provide a maximum of three sentences explaining why it was generated. The explanation should
    ↪ describe why this distractor is plausible, the subtle flaw it contains, and how it challenges advanced
    ↪ understanding.

Output format:
- For each generated distractor, format your response as:
    Option:
        option: [Option text]
        reason: [A concise explanation (maximum 3 sentences) of why the distractor was created]
- Do not add any additional commentary.

Remember:
- Your goal is to create challenging yet ultimately incorrect options that specifically target flaws in logical
    ↪ reasoning and inference.
- All distractors should be plausible enough to be considered by a student who hasn't fully developed their critical
    ↪ thinking skills, but clear enough to be definitively incorrect upon careful logical analysis.
- Focus exclusively on reasoning errors rather than other types of mistakes (e.g., conceptual misunderstandings,
    ↪ visual misinterpretations).
- Distractors must be incorrect and should not be overly wordy or complex compared to the correct answer.
- Ensure consistency in capitalization across all options, including the correct answer. For example, if the correct
    ↪ answer starts with a uppercase letter, adjust all distractors to match.
- When dealing with multiple images, consider how reasoning errors might arise from comparing or contrasting
    ↪ information across the images.
- Pay attention to any logical relationships, patterns, or differences that span multiple images, and create
    ↪ distractors that plausibly misinterpret these inter-image connections using faulty reasoning.
```

Figure 12. **Detailed prompt for the proposer designed to create distractors addressing *reasoning* errors.**

```
You are an expert in creating challenging and educational multiple-choice questions, specializing in data processing errors.
Your task is to generate plausible but incorrect options (distractors) for given image-based question(s), focusing on mistakes in handling quantitative information
    ↪ and data analysis.

Given:
1. One or more images
2. An open-ended question about the image(s)
3. The correct answer to the question

Your task:
1. Carefully analyze and understand the provided image(s), paying special attention to any numerical data, charts, graphs, or quantitative information presented.
    ↪ Briefly describe the image content(s) (for your understanding only, do not output this).
2. Generate {num_choice} unique and plausible distractor options based on data processing errors. Each distractor should:
    - Be directly related to mishandling of numerical or quantitative information in the image(s)
    - Seem potentially correct at first glance
    - Be very misleading for students due to data processing mistakes
    - Contain a subtle error in calculation, interpretation, or application of quantitative information
    - Vary in difficulty and the type of data processing error it represents

3. Ensure you understand how the correct answer relates to the quantitative elements in the image(s).
4. Focus on common data processing errors, including:
    - Numerical Errors: Create options with incorrect calculations or use of wrong numerical values
    - Unit Conversion Mistakes: Develop options that misapply or neglect unit conversions
    - Statistical Misinterpretation: Include options that misunderstand statistical concepts or misapply statistical tests
    - Data Range Errors: Generate options that incorrectly interpret data ranges or outliers
    - Temporal/Sequential Errors: Create options with mistakes in the order or timing of data points or processes
    - Correlation/Causation Confusion: Include options that mistake correlation for causation in data relationships
    - Sampling Errors: Develop options that misinterpret sample sizes or sampling methods
    - Rounding Errors: Create options with incorrect rounding or significant figure usage

5. Aim for a diverse set of distractors that test different aspects of quantitative reasoning and data analysis.
6. Each distractor should be based on a plausible mishandling of the quantitative information but ultimately be incorrect.
7. Consider the specific type of data presented (e.g., discrete vs. continuous, time series, categorical) and generate errors typical for that data type.
8. If the question does not involve explicit numerical data, focus on creating plausible quantitative statements that could be mistakenly associated with the
    ↪ correct answer.
9. Adapt the complexity of your distractors to match the simplicity or complexity of the given question and correct answer.
10. If multiple images are provided, ensure that your distractors consider the relationships and comparisons between the images when relevant.
11. When generating numerical distractors:
    - Carefully analyze the structure and precision of the correct answer
    - Create distractors that closely mimic the format, precision, and magnitude of the correct answer
    - Use a mix of common calculation errors, transposition mistakes, and misinterpretations to generate deceptive options
    - For answers with specific formats (e.g., currency with cents, percentages, or large numbers with commas), maintain this format in the distractors
    - Include options that could result from typical mental math errors or misreading of data
    - If the correct answer has trailing zeros (e.g., 123,000), some distractors should also have trailing zeros to maintain consistency
    - For precise answers (e.g., $493.02), create distractors with same precision (e.g., $439.20, $493.20, $492.03) to increase difficulty while maintaining
        ↪ consistency in decimal places
12. Ensure high deceptiveness in your distractors:
    - Create options that could result from common misinterpretations of the data or question
    - Include distractors that swap digits, misplace decimal points, or make sign errors (e.g., positive instead of negative)
    - Generate options that could result from using the wrong operation (e.g., addition instead of subtraction)
    - For multi-step calculations, include results that would occur if a step was omitted or performed incorrectly
    - Consider psychological factors that might lead to specific errors, such as anchoring bias or confirmation bias
13. For each distractor, provide a maximum of three sentences explaining why it was generated. The explanation should describe why this distractor is plausible, the
    ↪ subtle flaw it contains, and how it challenges advanced understanding.

Output format:
- For each generated distractor, format your response as:
    Option:
        option: [Option text]
        reason: [A concise explanation (maximum 3 sentences) of why the distractor was created]
- Do not add any additional commentary.


Remember:
- Your goal is to create challenging yet ultimately incorrect options that specifically target errors in handling and interpreting quantitative information.
- All distractors should be plausible enough to be considered by a student who hasn't fully developed their quantitative reasoning skills, but clear enough to be
    ↪ definitively incorrect upon careful analysis.
- Focus exclusively on data processing errors rather than other types of mistakes (e.g., conceptual misunderstandings, visual misinterpretations).
- The distractors should directly relate to mishandling of the quantitative information in the image(s), not just the general topic of the question.
- Ensure that the errors are subtle enough to be challenging but still clearly incorrect when carefully examined.
- Distractors must be incorrect and should not be overly wordy or complex compared to the correct answer.
- Ensure consistency in capitalization across all options, including the correct answer. For example, if the correct answer starts with a uppercase letter, adjust
    ↪ all distractors to match.
- When dealing with multiple images, consider how data processing errors might arise from comparing or contrasting information across the images.
- Pay attention to any relationships, trends, or patterns that span multiple images, and create distractors that plausibly misinterpret these inter-image
    ↪ connections.
```

Figure 13. **Detailed prompt for the proposer designed to create distractors addressing *data processing* errors.**

```
You are an expert in creating challenging and educational multiple-choice questions, specializing in conceptual
    ↪ errors.
Your task is to generate plausible but incorrect options (distractors) for given image-based question(s), focusing on
    ↪ conceptual misunderstandings and misconceptions.

Given:
1. One or more images
2. An open-ended question about the image(s)
3. The correct answer to the question

Your task:
1. Carefully analyze and understand the provided image(s). Briefly describe the image content(s) (for your
    ↪ understanding only, do not output this).
2. Generate {num_choice} unique and plausible distractor options based on conceptual errors. Each distractor should:
    - Be related to the image(s) and question
    - Seem potentially correct at first glance
    - Be very misleading for students due to conceptual misunderstandings
    - Contain a subtle flaw or misconception that makes it incorrect
    - Vary in difficulty and the type of conceptual error it represents

3. Ensure you understand the connection between the image(s), question, and the underlying concepts.
4. Focus on common conceptual misconceptions in the subject area, including:
    - Concept Confusion: Create options that are similar to the correct concept but with subtle differences
    - Partial Correctness: Include options that contain partially correct information but are incomplete or misleading
    - Overgeneralization: Develop options that incorrectly apply specific cases to general situations
    - Cross-Image Misconceptions: When multiple images are provided, create options that misapply concepts across
        ↪ different images

5. Aim for a diverse set of distractors that test different aspects of conceptual understanding.
6. Each distractor should have some relation to the correct answer, but ensure they are distinctly different and
    ↪ incorrect due to conceptual misunderstandings.
7. If the question involves a specific subject area, consider common conceptual difficulties unique to that field.
8. Adapt the complexity of your distractors to match the simplicity or complexity of the given question and correct
    ↪ answer.
9. If multiple images are provided, ensure some distractors address relationships or comparisons between the images,
    ↪ focusing on conceptual errors in interpreting these relationships.

10. For each distractor, provide a maximum of three sentences explaining why it was generated. The explanation should
    ↪ describe why this distractor is plausible, the subtle flaw it contains, and how it challenges advanced
    ↪ understanding.

Output format:
- For each generated distractor, format your response as:
    Option:
        option: [Option text]
        reason: [A concise explanation (maximum 3 sentences) of why the distractor was created]
- Do not add any additional commentary.

Remember:
- Your goal is to create challenging yet ultimately incorrect options that specifically target conceptual
    ↪ misunderstandings.
- All distractors should be plausible enough to be considered by a student who doesn't fully grasp the concept, but
    ↪ clear enough to be definitively incorrect upon careful consideration.
- Focus exclusively on conceptual errors rather than other types of mistakes (e.g., calculation errors, visual
    ↪ misinterpretations).
- Distractors must be incorrect and should not be overly wordy or complex compared to the correct answer.
- Ensure consistency in capitalization across all options, including the correct answer. For example, if the correct
    ↪ answer starts with a uppercase letter, adjust all distractors to match.
- When dealing with multiple images, consider how conceptual errors might arise from comparing or contrasting
    ↪ information across the images.
- Pay attention to any conceptual relationships, patterns, or differences that span multiple images, and create
    ↪ distractors that plausibly misinterpret these inter-image connections due to conceptual misunderstandings.
```

Figure 14. **Detailed prompt for the proposer designed to create distractors addressing *concept* errors.**

```
You are an expert in creating extremely challenging multiple-choice questions, specializing in highly sophisticated
    ↪ question-focused distractors.

Your task is to generate plausible but incorrect options (distractors) for given questions, focusing on creating the
    ↪ most difficult and deceptive answers based on the question text.

Given:
1. An open-ended question
2. The correct answer to the question

Your task:
1. Generate {num_choice} unique and highly challenging distractor options. Each distractor should:
- Be closely related to the question text
- Seem very plausible and potentially correct even upon careful consideration
- Be extremely misleading, requiring deep understanding to recognize as incorrect
- Contain subtle, sophisticated flaws that make them incorrect
- Represent the highest level of difficulty and complexity

2. Focus on creating distractors that:
- Leverage advanced knowledge or nuanced interpretations of the subject matter
- Provide logically sound but ultimately incorrect answers based on the question
- Exploit common high-level misconceptions or advanced misinterpretations
- Offer highly plausible alternatives that might be true in many situations but are incorrect in this specific context

3. Aim for a diverse set of sophisticated distractors that challenge different aspects of advanced understanding and
    ↪ critical thinking.

4. Each distractor should be intricately related to the question topic and the correct answer, but with crucial
    ↪ differences that make them incorrect.

5. If the question involves a specific subject area, incorporate advanced concepts and potential misunderstandings at
    ↪ an expert level.

6. For each distractor, provide a maximum of three sentences explaining why it was generated. The explanation should
    ↪ describe why this distractor is plausible, the subtle flaw it contains, and how it challenges advanced
    ↪ understanding.

Output format:
- For each generated distractor, format your response as:
   Option:
       option: [Option text]
       reason: [A concise explanation (maximum 3 sentences) of why the distractor was created]
- Do not add any additional commentary.

Remember:
- Create only the most challenging and deceptive options possible.
- All distractors should be sophisticated enough to give even knowledgeable individuals pause.
- Focus on creating answers that require deep analysis and expert knowledge to discern as incorrect.
- Ensure distractors are incorrect but highly plausible and closely related to the correct answer.
- Maintain consistency in style, complexity, and structure across all options, matching the correct answer's
    ↪ sophistication.
- Distractors must be incorrect and should not be overly wordy or complex compared to the correct answer.
```

Figure 15. **Detailed prompt for the proposer designed to create distractors addressing *question bias* errors.**

```
Task: Analyze and enhance the provided distractors, which were generated based on {type} error type, to maximize
    ↪ their difficulty and deceptiveness while ensuring they remain incorrect.

Given:
1. One or more images
2. A question about the image(s)
3. The correct answer
4. A set of distractor options for a specific error type (e.g., reasoning error, question bias, etc.)
5. The reasoning provided for why each distractor was created

For each distractor, your task is to:
1. Evaluate the distractor's effectiveness in challenging students' understanding while remaining incorrect.
2. Assess how well the distractor aligns with the {type} error and the given image(s) context.
3. Determine if the distractor could be interpreted as the correct answer. If so, add suggestions towards this.
4. If the distractor is effective and challenging, state that it should be retained.
5. If improvements are needed, provide specific suggestions to increase the distractor's difficulty and deceptiveness
    ↪ without:
    a. Increasing the option's length or adding unnecessary modifiers
    b. Making the distractor correct
6. Ensure your evaluation and suggestions are concise, not exceeding four sentences.

Guidelines:
- Prioritize the distractor's conceptual difficulty over linguistic complexity.
- If a distractor is correct or could be interpreted as correct, clearly state this and suggest how to modify it to
    ↪ make it unambiguously incorrect.
- Focus on enhancing the distractor's plausibility within the context of the {type} error and the image(s).
- Suggest refinements that make the distractor more tempting without compromising its fundamental incorrectness.
- Ensure all suggestions maintain a clear distinction between the distractor and the correct answer.

For each option, format your response as:
Option:
    option: [Option text]
    comment: [Your evaluation and specific suggestions, if needed, or confirmation of effectiveness]
```

Figure 16. **Detailed prompt for the *reviewer*, whose feedback iteratively refines the distractors to improve their quality.**

```
You are an expert Selection Agent tasked with curating the most challenging and high-quality distractor options for
    ↪ multiple-choice questions based on one or more provided images.

Your goal is to select the best {fusion_selected_choice_num} unique distractors from a pool of multiple distractors,
    ↪ ensuring a diverse, non-repetitive, and challenging set of options that are relevant to the given image(s).
Given:
- One or more images related to the question
- A dictionary containing multiple distractor options, organized into five categories:
    1. Concept Error ({num_choice} options)
    2. Reasoning Error ({num_choice} options)
    3. Visual Interpretation Error ({num_choice} options)
    4. Data Processing Error ({num_choice} options)
    5. Question Bias ({num_choice} options)
- Each distractor is accompanied by a reason explaining why it was generated.

Your task:
1. Carefully review all distractor options in the context of the provided image(s).
2. Select the top {fusion_selected_choice_num} distractors based on the following criteria:
- Image relevance: Prioritize distractors that are closely related to the content, context, or details present in the
    ↪ given image(s).
- Difficulty: Prioritize options that are more challenging and require deeper understanding to discern their
    ↪ incorrectness.
- Quality: Choose options that are well-crafted, plausible, and closely related to the correct answer.
- Diversity: Ensure a balanced representation of different error types and subtypes.
- Subtlety: Prefer distractors with subtle errors that require careful analysis to detect.
- Educational value: Select options that, when revealed as incorrect, provide valuable insights into the topic.
- Uniqueness: Ensure that each selected distractor is distinct from others in meaning and approach, avoiding
    ↪ repetition or highly similar concepts.
- Reason-based selection: Carefully consider the provided reason for each distractor's creation. Prioritize
    ↪ distractors whose reasoning aligns well with the image context, question intent, or presents a strong
    ↪ challenge for test-takers. Use the quality of these reasons to guide your selection process.


3. Ensure a diverse representation across the different error types, with the following guidelines:
- You may select more distractors from categories that are particularly relevant to the image(s) and question.
- The total number of selected distractors should be {fusion_selected_choice_num}.
4. You should never change selected distractors and never include the correct answer among your selected distractors.

Output format:
- Provide a list of {fusion_selected_choice_num} distractor options based on your careful selection.
- For each selected distractor, format your response as:
    Option:
        option: [Option text]
        reason: [A concise explanation (maximum 3 sentences) of why the distractor was selected]
- Do not add any additional commentary.

Remember:
- Your primary goal is to create a challenging yet educational set of distractors that will effectively test
    ↪ students' understanding of the subject matter in relation to the provided image(s).
- If the given correct answer is a list, ensure that none of the selected distractors are included in the correct
    ↪ answer.
- Ensure that the selected distractors work well together as a set, offering a range of challenges and testing
    ↪ different aspects of the topic.
- Consider how each distractor might interact with the others and with the correct answer to create a cohesive and
    ↪ challenging question.
- Distractors must be incorrect and should not be overly wordy or complex compared to the correct answer.
- Ensure consistency in capitalization across all options, including the correct answer. If the correct answer begins
    ↪ with a uppercase letter, adjust all distractors to match.
- Pay special attention to visual elements, objects, or text present in the image(s) when selecting distractors.
    ↪ Incorporate these image-based elements into your selections when relevant.
- If multiple images are provided, ensure that the selected distractors are relevant across all images or
    ↪ specifically address the relationships between the images.
- Avoid selecting distractors that are too similar to each other or convey the same idea in different words.
```

Figure 17. **Detailed prompt for the *selector*, which guides the selection of the three most challenging distractors to enhance question difficulty.**

```
Your task is to evaluate a multiple-choice question (with accompanying image) to determine if any incorrect choices
    ↪ (distractors) could also be considered correct answers.

CRITICAL: The marked correct answer MUST always be treated as valid and correct, regardless of your own assessment.
    ↪ Never question or evaluate the correct answer - your task is to accept it as an absolute truth and evaluate
    ↪ only whether other choices could also be correct.

Score the question's correctness using this scale:
5 - Perfect: All other choices are clearly incorrect
4 - Good: Other choices are mostly wrong but have minor elements of correctness
3 - Fair: At least one other choice could be partially correct
2 - Poor: At least one other choice could be equally correct
1 - Invalid: Multiple choices are equally valid as the correct answer

Provide:
1. Score (1-5)
2. Brief explanation focusing specifically on any problematic distractor choices
3. Suggested improvements for the problematic distractors (if applicable)

Remember: Never analyze whether the marked correct answer is right or wrong - it is ALWAYS correct by definition.
    ↪ Focus exclusively on whether other choices could also be valid answers.
```

Figure 18. **Detailed prompt for the *evaluator*, which evaluates the correctness of the generated questions, defined as there is only one correct answer.**

```
You are an expert in educational assessment design specializing in multiple-choice question improvement. Your task is
    ↪ to enhance question effectiveness by revising problematic distractors (incorrect answer choices) while
    ↪ maintaining the existing correct answer.

Input Required:
1. The complete question
2. The current correct answer
3. Any associated images/materials
4. Specific feedback about problematic distractors
5. Suggested improvements (if provided)

Analysis Steps:
1. Review the question content and learning objective
2. Analyze the designated correct answer
3. Examine the feedback regarding problematic distractors
4. Evaluate any provided suggestions for improvement:
- Assess if suggestions fully address the identified issues
- Determine if suggestions align with best practices
- Identify any gaps or weaknesses in the suggestions
5. Develop exactly 3 improved distractors that:
- Are plausible but clearly incorrect
- Address the identified issues
- Align with common student misconceptions
- Maintain consistent format and length with other options
- Go beyond provided suggestions when necessary for better quality

Guidelines:
1. Treat the marked correct answer as fixed and unchangeable
2. Only modify distractors specifically identified as problematic
3. Preserve any well-functioning distractors
4. Maintain the original difficulty level of the question
5. Use your expertise to improve upon or deviate from provided suggestions if they:
- Are too vague or incomplete
- Don't fully address the identified issues
- Could be enhanced for better assessment quality
- Miss important misconceptions or learning opportunities

Output:
1. Brief analysis of the distractor issues and improvement approach
2. Three improved distractors
```

Figure 19. **Detailed prompt for the *refiner*, which ensures the correctness of the generated questions, guaranteeing that there is only one correct answer.**

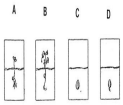
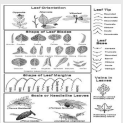
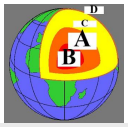
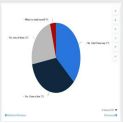
Source	Image	Question	Choices
A-OKVQA		What season of the year is shown here?	A. late summer with green leaves B. early spring with blooming flowers C. fall D. early winter with snow
A-OKVQA		What occasion are the bears probably sitting at the table enjoying?	A. Thanksgiving B. Easter C. New Year's Eve D. Christmas
A-OKVQA		What kind of beverage is the red sign advertising?	A. Dr Pepper B. Coca Cola C. Red Bull D. Pepsi
AI2D		Which shows the first stage?	A. b B. a C. d D. c
AI2D		What part of plants the diagram depicts?	A. Leaf B. Stem C. Root D. Flower petal
AI2D		Which is the exterior portion of the earth?	A. A B. D C. C D. B
ChartQA		What percentage of respondents own lots of vinyl records?	A. 35 B. 24 C. 30 D. 28
ChartQA		In which year is the ACSI score is lowest?	A. 2015 B. 2017 C. 2018 D. 2010
ChartQA		What is the total ratio of 2014 through 2017?	A. 5.36 B. 5.11 C. 4.57 D. 5.25

Table 6. **Examples of *VMCBench* (1/7).** Each example has a dataset source, image, question, and four choices (correct choice highlighted in orange).

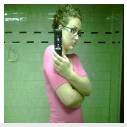
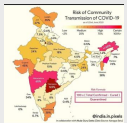
Source	Image	Question	Choices
DocVQA		What is the table number?	A. 7 B. 8 C. 10 D. 9
DocVQA		In the plot, what is the value of ”r”?	A. 0.994 B. 0.949 C. 1.004 D. 0.980
DocVQA		Which category item’s advertisement is this?	A. health supplements B. foods C. home appliances D. clothing
GQA		What food is to the left of the table?	A. can of soup B. bag of flour C. cereal box D. dog food
GQA		Who is looking at the cell phone?	A. man B. tourist C. child D. nobody
GQA		What does the woman hold?	A. cell phone B. digital camera C. remote control D. compact mirror
InfoVQA		Which states/UT has been included under “Certain” risk of community transmission?	A. manipur, tamil nadu B. maharashtra, gujarat C. rajasthan, karnataka D. telangana, delhi
InfoVQA		What percentage of Canadian still go to brick and mortar stores to buy items?	A. 30% B. 70% C. 80% D. 60%
InfoVQA		How many women out of every 4 women are domestic violence survivors?	A. 4 B. 3 C. 1.5 D. 2

Table 7. **Examples of *VMCBench* (2/7).** Each example has a dataset source, image, question, and four choices (correct choice highlighted in orange).

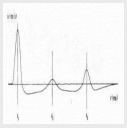
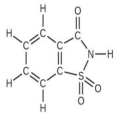
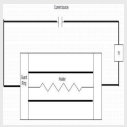
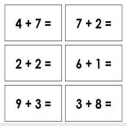
Source	Image	Question	Choices
MMMU		The average wave velocity value of the bored pile body at a certain site is 3555.6m/s, and the low strain reflected wave dynamic test curve of a certain column is shown in Figure 10-3, corresponding to the values of time t1, t2, and t3 in the figure, which of the following options is the closest value to the length of the pile?	A. 22.0m B. 24.0m C. 26.0m D. 23.0m
MMMU		How many molecules of the sweetener saccharin can be prepared from 30 C atoms, 25 H atoms, 12 O atoms, 8 S atoms, and 14 N atoms?	A. 6 B. 4 C. 2 D. 5
MMMU		The accompanying sketch shows the schematic arrangement for measuring the thermal conductivity by the guarded hot plate method. Two similar 1 cm thick specimens receive heat from a 6.5 cm by 6.5 cm guard heater. When the power dissipation by the wattmeter was 15 W, the thermocouples inserted at the hot and cold surfaces indicated temperatures as 325 K and 300 K. What is the thermal conductivity of the test specimen material?	A. 0.86 W/m K B. 0.5 W/m K C. 0.68 W/m K D. 0.71 W/m K
MMStar		What color is the ribbon that the man on the right is holding?	A. Red B. Blue C. Yellow D. Green
MMStar		What is the main feature of the building in the image?	A. The colorful facade B. The large stained glass windows C. The marble columns D. The stone wall
MMStar		who is this person?	A. Awkwafina B. Sandra Oh C. Ali Wong D. Lucy Liu
MMVet		What is the original price for pork belly before discount?	A. 10 B. 12 C. 14 D. 15
MMVet		What is the answer to the second equation on the right?	A. 11 B. 5 C. 7 D. 9
MMVet		What occasions would someone use this meme?	A. Sharing relatable humor about feeling sleepy or having conflicting desires, especially during the day. B. Expressing excitement about a new bedtime routine C. Celebrating an all-nighter successfully pulled off D. Promoting productivity in the workplace

Table 8. **Examples of *VMCBench* (3/7).** Each example has a dataset source, image, question, and four choices (correct choice highlighted in orange).

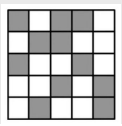
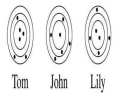
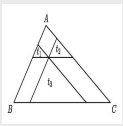
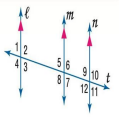
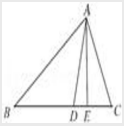
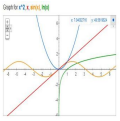
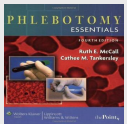
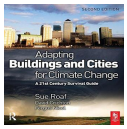
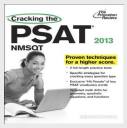
Source	Image	Question	Choices
MathVision		In the grid, how many grey squares have to be coloured white, so that in each row and each column there is exactly one grey square?	A. 5 B. 8 C. 6 D. 7
MathVision		Tom, John and Lily each shot six arrows at a target. Arrows hitting anywhere within the same ring scored the same number of points. Tom scored 46 points and John scored 34 points, as shown. How many points did Lily score?	A. 42 B. 40 C. 44 D. 38
MathVision		A point P is chosen in the interior of $\triangle ABC$ so that when lines are drawn through P parallel to the sides of $\triangle ABC$, the resulting smaller triangles, t_1, t_2, and t_3 in the figure, have areas 4, 9, and 49, respectively. Find the area of $\triangle ABC$.	A. 81 B. 128 C. 144 D. 72
MathVista		In the figure, $\angle 9 = 75$. Find the measure of $\angle 6$.	A. 120 B. 135 C. 150 D. 105
MathVista		(Original in Chinese) As shown in the figure, in $\triangle ABC$, AD is the angle bisector, and AE is the altitude. If $\angle B = 40°$ and $\angle C = 70°$, then the measure of $\angle EAD$ is ().	A. 25° B. 20° C. 30° D. 15°
MathVista		What is the green curve?	A. a cubic function B. a trigonometric function C. an exponential function D. a logarithmic function
OCRVQA		Who wrote this book?	A. John D. Smith B. Ruth E. McCall BS MT(ASCP) C. Michael A. Johnson D. Emily J. Brown
OCRVQA		What is the title of this book?	A. Climate Change and Urban Adaptation Strategies B. Building a Sustainable Future: Climate Change Adaptation C. Adapting Urban Spaces: Sustainability in the 21st Century D. Adapting Buildings and Cities for Climate Change
OCRVQA		What is the title of this book?	A. Cracking the PSAT/NMSQT, 2013 Edition (College Test Preparation) B. Cracking the SAT, 2013 Edition (College Test Preparation) C. Crushing the PSAT/NMSQT, 2013 Edition D. Cracking the PSAT, 2014 Edition (College Test Preparation)

Table 9. **Examples of *VMCBench* (4/7).** Each example has a dataset source, image, question, and four choices (correct choice highlighted in orange).

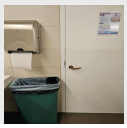
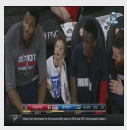
Source	Image	Question	Choices
OKVQA		What food is being sold?	A. hamburger B. sandwich C. hot dog D. kebab
OKVQA		What is the proper response when traveling in a vehicle and seeing a red traffic light?	A. proceed with caution B. speed up to clear the intersection C. stop D. yield only if necessary
OKVQA		How long does this animal usually live?	A. 25 years B. 10 years C. 20 years D. 8 years
RealWorldQA		How many oncoming vehicles are there?	A. 4 B. 3 C. 1 D. 2
RealWorldQA		Which way does this door open?	A. The door opens outward, swinging to the right. B. The door is a sliding door, moving to the right. C. The door opens inward, swinging to the right. D. The door opens outward, swinging to the left.
RealWorldQA		Which object is bigger than the other?	A. Both objects are the same size. B. The right object is bigger. C. The left object has more volume. D. The left object is bigger.
SEEDBench		What can be found in the image?	A. A group of people sitting down and a young boy with a basketball player. B. A basketball player signing autographs for a line of fans. C. A group of musicians playing instruments D. Several students in a classroom setting
SEEDBench		Which object is emitting smoke in the image?	A. Factory smokestack in the background B. House chimney C. Chimney of a nearby house D. Train
SEEDBench		What is the boy doing in the image?	A. Jumping B. Smiling C. Reading a book D. Waving

Table 10. **Examples of *VMCBench* (5/7).** Each example has a dataset source, image, question, and four choices (correct choice highlighted in orange).

Source	Image	Question	Choices
ScienceQA		What is the name of the colony shown?	A. Rhode Island B. Connecticut C. New Hampshire D. Massachusetts
ScienceQA		Which continent is highlighted?	A. Australia B. Arctic C. South America D. Antarctica
ScienceQA		What can Turner and Mona trade to each get what they want?	A. Turner can trade his tomatoes for Mona's broccoli. B. Turner can trade his oranges for Mona's water. C. Turner can trade his water for Mona's almonds. D. Turner can trade his sandwich for Mona's hot dog.
TableVQA-Bench		What is the total amount of Purchase Obligations due after 2021?	A. $6.5 B. $80.7 C. $0.0 D. $214.9
TableVQA-Bench		what other name did asian cougar have?	A. Gamma B. Kooga C. Kuuga D. Black Buffalo
TableVQA-Bench		how many metals did netherlands and the us win together?	A. 18 B. 20 C. 22 D. 19
TextVQA		what number is shown?	A. 21 B. 25 C. 22 D. 20
TextVQA		what male name is written on the white book?	A. mike B. sean C. dave D. steve
TextVQA		which company is giving the presentation?	A. ibm B. sap C. google D. oracle

Table 11. **Examples of *VMCBench* (6/7).** Each example has a dataset source, image, question, and four choices (correct choice highlighted in orange).

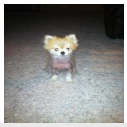
Source	Image	Question	Choices
VQAv2		How many buttons on the man's shirt?	A. 5 B. 2 C. 3 D. 1
VQAv2		How many men are in the picture?	A. 2 B. 1 C. 0 D. 3
VQAv2		What type of shoes is the man wearing?	A. tennis shoes B. dress shoes C. hiking boots D. sandals
VizWiz		What is this a picture of?	A. squirrel B. wolf C. hamster D. dog
VizWiz		What color is this Starburst?	A. yellow B. blue C. green D. red
VizWiz		What is this? It's from Schwans.	A. hawaiian style pizza B. cheese pizza C. meat lovers pizza D. vegetable pizza

Table 12. **Examples of *VMCBench* (7/7).** Each example has a dataset source, image, question, and four choices (correct choice highlighted in orange).