

A. Overview of the Appendix

- Section B illustrates the differences between Multi-Crit and prior multimodal judge benchmarks.
- Section C provides additional benchmark details, including the response models, evaluation criteria, judge prompt template, and supplementary data statistics.
- Section D presents qualitative examples of multimodal prompts and complete evaluation instances.
- Section E reports additional results and analysis.

B. Comparison Against Prior Benchmarks

As shown in Table 5, earlier benchmarks assign only a single overall preference label for each pair of model responses, whereas Multi-Crit provides multiple criterion-level human annotations. This makes Multi-Crit the first benchmark capable of assessing whether LMM judges can follow pluralistic evaluation criteria and deliver reliable criterion-level judgments. In addition, the average length of candidate model responses in Multi-Crit is significantly longer (164 vs. 99 words), further highlighting the increased difficulty and overall challenge of the benchmark.

Benchmark	Resp. Quantile	Multi-Criterion	Anno.
MLLM-as-a-Judge [3]	(54, 89, 153)		✗
VL-RewardBench [23]	(48, 99, 136)		✗
MM-RLHF [68]	(16, 49, 144)		✗
Multimodal-RewardBench [56]	(1, 72, 138)		✗
Multi-Crit (ours)	(104, 164, 247)		✓

Table 5. Comparison with existing benchmarks.

C. Benchmark Construction and Statistics

C.1. Additional Benchmark Details

Models for Generating Candidate Responses: Gemini-2.5-Pro [8], Gemini-2.5-Flash [8], GPT-4.1 [30], GPT-4o [16], Qwen2.5-VL-7B-Instruct [2], Qwen2.5-VL-72B-Instruct [2], LLaMA-3.2-11B-Vision-Instruct [11], InternVL3-8B-Instruct [73], InternVL3-38B-Instruct [73], MiMo-VL-7B-RL [39], GLM-4.1V-9B-Thinking [14].

Specifically, for prompts from WildVision-Battle [27], if the two original responses receive a tie in their human preference, we keep the original pair; otherwise, we generate new responses using the models listed above.

Detailed Evaluation Criteria. Table 7 presents the detailed criteria for judging open-ended generation, and Table 8 lists those for judging verifiable reasoning. To better guide both human annotators and judge LMMs, we organize the description of each criterion into three components:

1. **Definition:** A general descriptions of the criterion and the key aspects a response should demonstrate.

2. **Positive Indicators:** Behaviors and qualities that are encouraged and should be preferred in responses.
3. **Negative Indicators:** Errors, undesirable behaviors, or deficiencies that should be penalized under this criterion.

Evaluation Prompt Template. We provide the evaluation prompt used for judge-model inference in Table 9. When evaluating LMM judges on the Multi-Crit benchmark, each criterion instance—consisting of the question, the pair of responses, and the target criterion—is filled into this template and sent to the model during inference. For finetuned judges [45, 48, 53, 66], we adapt this template to match the prompt format used in their original repositories, avoiding significant out-of-distribution judgment behaviors.

C.2. Additional Data Statistics

Response Pair Composition. As shown in Table 6, for both splits, roughly 40% of response pairs originate from the same model via random sampling, while the remaining pairs are generated by different models. These two types serve complementary roles: cross-model pairs reflect realistic deployment scenarios for automated evaluation (e.g., LMArena or WildVision Arena), where responses differ along multiple dimensions; intra-model pairs control for model identity and surface style, which is crucial when applying judges for test-time scaling and RL reward modeling. Our balanced mix ensures that Multi-Crit assesses whether LMM judges can capture quality differences both across models (reflecting model diversity) and within the same model (reflecting sampling diversity).

Split	#Prompts	Cross-model pairs	Intra-model pairs
Open-ended	299	168 (56.2%)	131 (43.8%)
Reasoning	126	76 (60.3%)	50 (39.7%)
Total	425	244 (57.4%)	181 (42.6%)

Table 6. Composition of response pairs.

Cross-Criterion Correlation and Conflict Frequency.

We report additional statistics on criterion-level human preferences to illustrate how Multi-Crit captures cross-criterion correlations and conflicts. Figure 8 shows how human preferences correlate across criteria on prompts that exhibit criterion-level conflicts. Figure 9 reports how often two criteria yield opposite human preferences on the same response pair. Multi-Crit contains preference conflicts across all criterion pairs in both splits, reflecting both the nuanced differences within the response pairs and the inherent diversity and tension across evaluation dimensions. Certain criterion pairs display intuitive trade-offs—such as *completeness vs. no-hallucination* and *no-hallucination vs. expressiveness* in the open-ended split, and *exploration vs. efficiency* in the reasoning split—and these pairs also exhibit the highest frequencies of human-annotated conflicts.

Evaluation Criterion	Description
Completeness and Coverage	<p>The response should provide a thorough and well-developed answer that fully addresses the intent of the prompt. It must address the complete scope of the task, incorporating all major elements specified in the prompt, as well as relevant visual aspects and broader contextual cues. When appropriate, drawing on relevant external knowledge is encouraged to enrich the explanation.</p> <ul style="list-style-type: none"> - Reward: The response addresses all key parts of the prompt and image, showing depth and effort in the description or explanation. - Penalize: The response is underdeveloped, fails to meet one or more specific requirements in the prompt, or omits important visual elements or interpretive points.
Visual Grounding and Details	<p>The response should demonstrate a clear and meaningful connection to the visual input. It should refer to observable elements in the image—such as objects, spatial relationships, colors, or text—and build its description or interpretation based on those elements.</p> <ul style="list-style-type: none"> - Reward: The response explicitly references relevant visual details that are clearly visible in the image. - Penalize: The response fails to connect meaningfully to the image, or uses vague, generic language that lacks specific visual grounding.
Factuality / No Hallucination	<p>The response should avoid introducing any visual details, objects, relationships, or factual claims that are not present in the image or reasonably suggested by the prompt. This includes both visual hallucinations (e.g., describing elements not visible in the image) and factual inaccuracies in general knowledge.</p> <ul style="list-style-type: none"> - Reward: The response stays grounded in the image and prompt, without inventing visual elements or making unsupported factual claims. - Penalize: The response introduces hallucinated visual content or inaccurate factual statements that are unsupported or misleading.
Creativity and Expressiveness	<p>The response should show originality or stylistic flair for open-ended tasks, and knowledge-informed articulation with precision and depth for analytical tasks. All responses must remain contextually appropriate and grounded in the visual input, while enhancing richness, nuance, and overall engagement.</p> <ul style="list-style-type: none"> - Reward: The response uses vivid language, unique phrasing, or inventive associations that enrich the interpretation, or it demonstrates professional articulation through deep and knowledge-grounded analysis. - Penalize: The response is overly literal, flat, or dull, lacking originality, variation in expression, or in analytical contexts, fails to demonstrate professional depth or expertise.
Clarity and Coherence	<p>The response should communicate ideas clearly and logically, with coherent structure and fluent language. This involves not only grammatical correctness, but also effective organization of information, smooth transitions, and consistent flow of ideas.</p> <ul style="list-style-type: none"> - Reward: The response is clearly written, logically structured, and easy to follow. A brief summary at the beginning may further improve clarity. - Penalize: The response is difficult to follow due to unclear structure, disorganized reasoning, poor transitions, or awkward and repetitive phrasing.

Table 7. Five evaluation criteria for judging *open-ended generation* tasks in Multi-Crit Benchmark.

Evaluation Criterion	Description
Visual Grounding	<p>The response should be explicitly grounded in the visual input. It must refer to salient visual elements—such as specific objects, spatial arrangements, colors, or visible text—and incorporate them meaningfully into the reasoning process. Visual references should be accurate and relevant to the task.</p> <ul style="list-style-type: none"> - Reward: The response clearly references important visual features and integrates them into the reasoning in a precise and relevant manner. - Penalize: The response fails to reference relevant visual elements, or uses generic or weakly connected visual details that do not meaningfully support the reasoning.
Logic Coherence and Consistency	<p>The reasoning should follow a logically sound and step-by-step progression, with each step building upon the previous one. The reasoning should be internally consistent, with no contradictions, missing steps, or unjustified leaps. The final answer should naturally and justifiably emerge from the reasoning process.</p> <ul style="list-style-type: none"> - Reward: The response presents a well-structured, internally consistent chain of reasoning that leads clearly and justifiably to the final answer. - Penalize: The response contains contradictions, missing steps, or disconnects between reasoning and answer. Short-cut behaviors—such as giving the final answer first with unsupported or inconsistent reasoning—should also be penalized.
Factuality / No Hallucination	<p>All claims and reasoning steps must be factually accurate and supported by the image or the prompt. The response should avoid hallucinated visual content, misidentifications, or factual inaccuracies in the reasoning process.</p> <ul style="list-style-type: none"> - Reward: The response is free from factual errors or hallucinations and relies only on valid observations and logical inferences. - Penalize: The response introduces hallucinated details, misidentifications, or incorrect factual claims that compromise the reasoning.
Reflection and Exploration	<p>The reasoning should demonstrate thoughtful reflection and a willingness to explore multiple possibilities, particularly when the task is ambiguous or complex. This includes acknowledging uncertainty, considering alternative interpretations, or revising initial assumptions.</p> <ul style="list-style-type: none"> - Reward: The response demonstrates depth through reflection, critical evaluation, or exploration of different solutions before reaching a conclusion. - Penalize: The response is overly rigid, superficial, or rushed, showing little to no depth of thought, reflection, or exploration of alternative possibilities.
Conciseness and Efficiency	<p>The reasoning should be clear, focused, and efficiently communicate the steps. It should avoid redundancy, digressions, or unnecessary elaboration that dilute the argument. For straightforward tasks, over-explaining or over-analyzing should also be avoided.</p> <ul style="list-style-type: none"> - Reward: The response is concise and well-structured, conveying reasoning steps precisely and proportionally to the task complexity. - Penalize: The response is verbose, repetitive, or includes irrelevant content that distracts from the reasoning. It may also overthink or over-explain simple prompts.

Table 8. Five evaluation criteria for judging *verifiable reasoning* tasks in Multi-Crit benchmark.

You are an expert in evaluating the quality of AI-generated responses according to specific evaluation criteria. Your task is to assess two responses generated by different AI assistants in reply to a user’s question about an image. The image is provided as part of the input. You must evaluate the responses **strictly and exclusively** based on the following evaluation criterion: {Criterion} Do not consider any other dimensions or criteria beyond what is specified above. Here are the inputs for your evaluation:
 [Question]: {Question}
 [Response 1]: {Response1}
 [Response 2]: {Response2}

First, provide a detailed justification for your evaluation. Refer to specific elements in the responses, how they align with the evaluation criterion, and relevant visual details from the image. On the final line, provide your final judgment on which response is better. Your judgment must be based solely on the specified criterion. Strictly follow this format: Response X is better.

Table 9. Evaluation prompt template used for LMM judge inference in Multi-Crit. Judge models are explicitly instructed to focus strictly and exclusively on the target evaluation criterion during each inference.

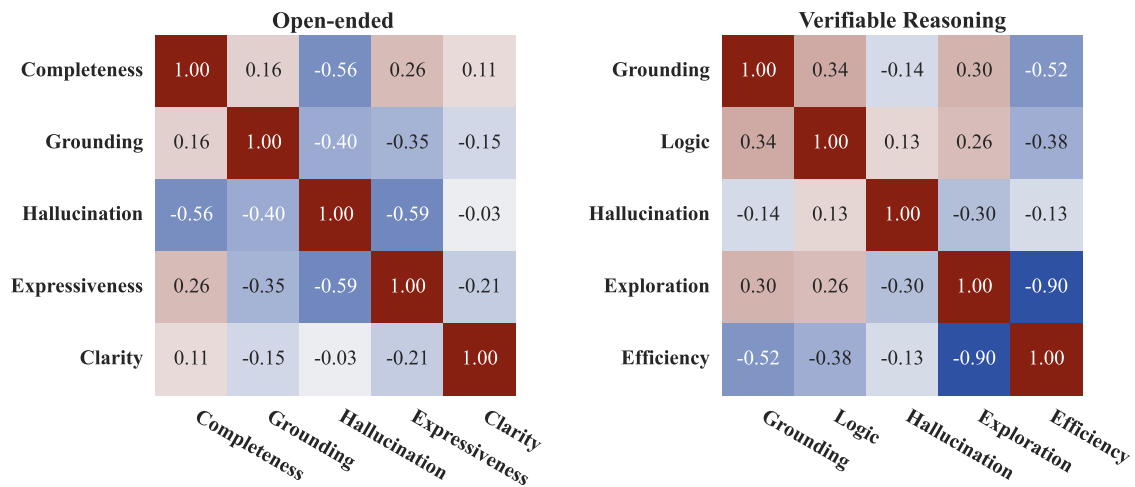


Figure 8. Correlation of criterion-level human preferences for prompts that exhibit preference conflicts in the open-ended (left) and reasoning (right) splits of Multi-Crit.

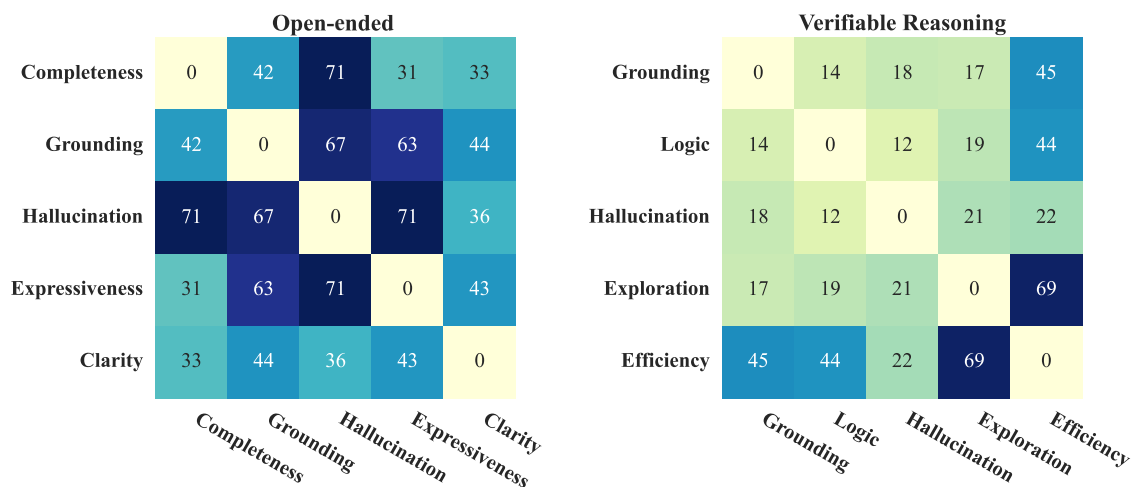


Figure 9. Counts of criterion pairs exhibiting human preference conflicts in the open-ended (left) and reasoning splits (right) of Multi-Crit.

D. Qualitative Examples

D.1. Multimodal Prompts

Figure 10 visualizes selected multimodal prompts from Multi-Crit, spanning diverse *open-ended generation* scenarios—such as captioning, storytelling, knowledge-based tasks, and text-rich understanding—as well as *verifiable reasoning* domains including math, science, coding, and visual analogy. These prompts cover a wide range of LMM-as-a-Judge use cases.

D.2. Full Evaluation Instances

We present six full evaluation instances—each consisting of a prompt and paired model responses—along with their multi-criterion human preference annotations. These examples illustrate how our curated data reflects fine-grained response differences and demonstrates criterion-level preference conflict patterns within each response pair.


The first three examples illustrate judgment in open-ended tasks:

- Table 10 shows an evaluation instance for judging two models in blind storytelling. This example reveals preference conflicts such as *completeness vs. grounding* and *grounding vs. no-hallucination*.
- Table 11 presents an evaluation instance for judging two models on captioning creative images. It exhibits representative preference conflicts including *completeness vs. no-hallucination*, *completeness vs. expressiveness*, and *expressiveness vs. clarity*.
- Table 12 provides an example of judging two responses from the same model on an instruction-rich image analysis task. Observed conflicts include *grounding vs. clarity* and the less common *no-hallucination vs. clarity*.

The following three examples illustrate judgment in verifiable reasoning tasks:


- Table 13 shows an evaluation instance for judging two models’ thinking traces on a math reasoning task. Although both traces lead to incorrect answers, one exhibits better *visual grounding* and *explores* alternative answers, whereas the other maintains stronger *logical consistency* with *fewer perceptual hallucinations*.
- Table 14 presents an evaluation instance for judging two thinking traces produced by the same model on object counting. While both responses lead to the correct final answer, one conducts *concise* reasoning, whereas the other identifies more *visual details* and actively *reflects on each object*, but introduces *minor hallucinations* that do not affect the final result.
- Table 15 provides an example of judging two models’ responses on a biological reasoning task involving a food web. In this case, one response is more *concise* and demonstrates deeper *reflection* by explicitly identifying a

key intermediate inference, while the other offers a more visually detailed description of the food-web structure.




Fine-grained Captioning

Q: Describe the image in detail.




Semiotics & Branding

Q: Based on this logo alone, what can you determine about the business it represents?




Action Recognition/ Comic

Q: This is a screenshot of my employee during his work time. What is he doing?




Blind Storytelling

Q: You are a storyteller for the blind. Please tell this comic book panel by naming character and all dialogues according to the image and the following script extract.



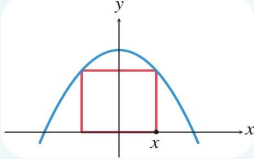
OCR & Document

Q: Extract the text and tabular data in the image. Create a json structure with bill contents and details like date name on the bill etc.,



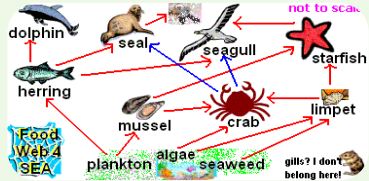
Geospatial World Knowledge

Q: What is the best way to commute from Trump Tower to the location shown in this image?



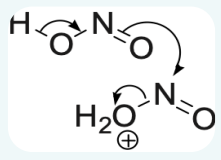
Math

Q: A rectangle is inscribed between $y=36-x^2$, as shown in the figure below. Write the area A of the rectangle as a function of x .
A: $72x-2x^3$



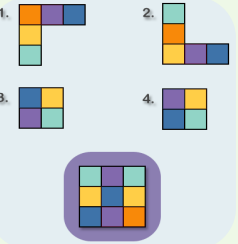
Biology

Q: The image shows a food web, where an arrow pointing from object A to object B indicates that A provides B with the energy needed for survival. According to the food web, what will happen if all the algae died? Choices: (A) Crabs and limpets will decrease (B) Dolphins will increase (C) Sea gulls will become extinct (D) Star fish will increase. **A:** A




Chemistry

Q: In the transition-state structure shown in the image, calculate the total number of bonds in the structure, including single, double, and triple bonds but excluding those involving hydrogen. **A:** 4




Visual Puzzles

Q: Which two puzzle pieces form the larger square? Options: (A) 1 & 2 (B) 1 & 3 (C) 1 & 4 (D) 2 & 3 (E) 2 & 4
A: B



Visual Analogy

Q: Based on the relationship between Origin and Analogy, select the option from the candidates that best completes this analogy. **A:** B



Coding

Q: Which code snippet below can possibly create the chart in the image? We are using Python version 3.11.0, matplotlib version 3.6.3, and seaborn version 0.12.2. (A) `import matplotlib.pyplot as plt...` (B) `...(C)...` (D) `...` **A:** B

Figure 10. Example prompts in the open-ended (top) and verifiable reasoning (bottom) splits of Multi-Crit. Our benchmark spans diverse multimodal prompts, reflecting the broad evaluation domains in which LMM judges are used.

Open-ended Example 1



Question You are a storyteller for the blind. Please tell this comic book panel by naming character according to the image and the following script extract. Corresponding character’s names have been added in white rectangles over the original image (do not mention rectangle or label in the storytelling).

DIALOGUE
 Marco: AN INVENTOR? I DON’T BELIEVE YOU!
 Babi: WHY NOT?

Response A (llava-v1.6-vicuna-7b) In the comic book panel, a character is named Marco. He is speaking to another character named Babi. Marco is questioning Babi’s claim that he is an inventor. Babi responds with an open-ended question, encouraging Marco to reconsider his skepticism. Marco’s skepticism towards Babi’s claim is evident from his words, “I DON’T BELIEVE YOU!” It seems like Marco is questioning Babi’s credibility or the authenticity of his claim. Babi’s open-ended question, “WHY NOT?”, indicates that he is not backing down and is pushing for a dialogue or a conversation where Marco can clarify his doubts. This comic panel seems to be a humorous take on skepticism and the importance of asking questions to explore different perspectives and possibilities.

Response B (yi-vl-plus) In a vividly colored comic book panel, we see two characters engaging in a lively conversation. One character, with a bright yellow head and a red shirt, exclaims, “An inventor? I don’t believe you!” His name, Marco, is displayed in a white rectangle. Opposite him, another character, sporting a black hairstyle and a blue shirt, confidently asks, “Why not?” His name, Babi, is also shown in a white rectangle. The contrasting colors of their clothes and hair add to the dynamic energy of their exchange, as they stand against a plain blue backdrop. The speech bubbles capture their dialogue, highlighting the playful and inquisitive spirit of their interaction.

Ground Truth	Completeness and Coverage	A > B
	Visual Grounding and Details	B > A
	Factuality / No Hallucination	A > B

Table 10. Example of an open-ended evaluation instance for blind storytelling. Response A conveys the moral of the story and fully follows the instruction to avoid mentioning rectangles. Response B includes more visual details but hallucinates the shirt colors.

Open-ended Example 2



Question

What does this image show?

Response A

(Llama-3.2-11B-Vision-Instruct)

The image depicts a surreal scene where a bottle of beer is pouring its contents onto the ground, creating a puddle that morphs into a miniature pool. Within this pool, two women are shown sitting and posing in swimsuits, as if they were enjoying a day at the beach or poolside. **The juxtaposition of the beer's spillage with the serene atmosphere of the women creates a humorous and unexpected visual contrast.**

Response B

(Qwen2.5-VL-7B-Instruct)

This image is a creative and surreal depiction that combines elements of everyday objects with imaginative scenarios. It shows a large, tilted bottle pouring liquid onto the ground. The liquid spreads out in a pool around two women who are sitting on the edge of the puddle. **One woman** is partially submerged in the liquid, while **the other sits on the edge looking at her.** The scene is contrasted by the dark, plain background, which makes the bottle and the women stand out prominently. This type of imagery is often used in **advertising or artistic representations** to convey a message or evoke emotions.

Ground Truth

Completeness and Coverage	B > A
Factuality / No Hallucination	A > B
Creativity and Expressiveness	A > B
Clarity and Coherence	B > A

Table 11. Example of an open-ended evaluation instance for creative-image captioning. Response A creatively highlights the **humorous and unexpected effect** and introduces **less hallucinated content**. Response B includes a brief **TL;DR** and an **inference** about image usage, with a more **coherent flow across elements**.

Open-ended Example 3



Question Create detailed one sentence image caption, capture the essence and include the following structured elements without assumptions or ambiguous language: **Background and Foreground:** Separate background objects and foreground objects to enhance focus; **Location:** The setting can be either indoors or outdoors; **Shot Size:** The image sizes: Extreme close up, Close up, Medium close up, Medium shot, Medium full shot, Full length. Use this parameter if a person is included in the photo; **Angle:** low angle shot, high angle shot, top view shot, flat lay. If the angle is straight, omit this specification; **Temperature:** Color temperature should be considered, with options including cold, warm, neutral, or vibrant. If the temperature is neutral, you may choose to not specify this detail; **Lens:** Use a fisheye lens or wide-angle as appropriate. If neither is used, this detail can be omitted; **Lighting:** Very important options include High-key Lighting, Low-key Lighting, Dramatic light, Backlight, silhouette, cinematic light; **Composition:** Employ techniques such as the rule of thirds, double exposure, over-the-shoulder shots, or symmetrical composition to structure the image; **Day Time:** Specify the time of day (e.g., blue hour, sunrise, morning, noon, afternoon) if this parameter is explicitly relevant; **Person Caption:** If a person is included, describe their race of skin color (Asian, Caucasian, Black, Arab, Indian, etc), hair (black, brown, blond, white/gray, and rarely red) view direction, emotions and mood, and clothing details.

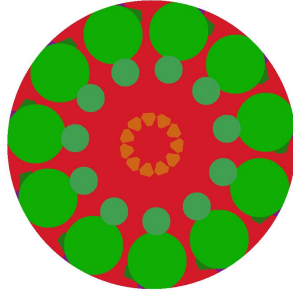
Response A (GPT-4.1) **Foreground:** Two children, one with light skin and straight brown hair, wearing a charcoal grey shirt and black shorts, and the other with dark skin and short curly black hair, dressed in a beige shirt and black shorts, are seated close to each other with the child on the right resting an arm around the other’s shoulders; both appear calm and friendly, looking directly at the camera. **Background:** Metal fencing with vertical green posts is visible, along with an overcast sky and faint outlines of a streetlight or utility pole. **Location:** Outdoors. **Shot Size:** Medium shot. **Angle:** Low angle shot. **Temperature:** Neutral. **Lighting:** Natural diffused daylight. **Composition:** Centered composition, subjects positioned side by side with upper bodies and faces as focal point. **Day Time:** Daytime.

Response B (GPT-4.1) **Foreground:** Two boys are sitting close together on what appears to be outdoor seating; **Background:** A wire mesh fence and holding poles are visible against a blank, overcast sky; **Location:** Outdoors; **Shot Size:** Medium shot; **Angle:** Low angle shot; **Temperature:** Neutral; **Lighting:** Diffused natural light due to an overcast day; **Composition:** Symmetrical composition with both boys centered in the frame; **Day Time:** Daytime, likely noon or afternoon given the soft daylight; **Person Caption:** On the left, a Caucasian boy with straight, shoulder-length brown hair faces the camera with a neutral, slightly guarded expression, wearing a charcoal grey henley shirt and black shorts; on the right, a Black boy with short, curly black hair also faces the camera with a calm, faintly smiling demeanor, dressed in a loose beige t-shirt and black shorts, his left arm draped supportively behind the other boy.

Ground Truth	Completeness and Coverage	B > A
	Visual Grounding and Details	A > B
	Factuality / No Hallucination	A > B
	Clarity and Coherence	B > A

Table 12. Example of an evaluation instance in instruction-rich image analysis. Both responses are randomly sampled from GPT-4.1. Response A provides more background details and is less hallucinated, while Response B better follows all required elements in the user prompt and offers a more clear and readable description of the foreground and composition.

Verifiable Reasoning Example 1



Question The entire pie seen in the figure is divided among several children. Each child receives a piece of pie, and each piece of pie looks identical. The maximum possible number of children there is: Choices: (A) 10 (B) 12 (C) 9 (D) 11.

Answer D

Response A (Qwen2.5-VL-72B-Instruct) To determine the maximum possible number of children who can receive an identical piece of the pie, we need to analyze the symmetry of the figure. 1) Identify the Symmetry: The figure appears to be a circular design with a central red area and green circles arranged around it. The key is to find the rotational symmetry of the design. 2) Count the green circles: There are 12 green circles arranged around the central red area. This suggests that the design might have 12-fold rotational symmetry. 3) If we rotate the figure by $\frac{360^\circ}{12} = 30^\circ$, the figure should look the same. This means that the design is symmetric under a 12-fold rotation. 4) Conclusion: Since the design has 12-fold rotational symmetry, the pie can be divided into 12 identical pieces. Therefore, the maximum possible number of children who can each receive an identical piece of the pie is 12. Thus, the correct answer is **B**.

Response B (Gemini-2.5-Flash) To find the maximum possible number of children, we need to determine the number of identical pieces the pie can be divided into. This means identifying the smallest repeating unit in the circular pattern. Let's observe the elements arranged in a circle: 1) Large green circles: Count the number of large green circles arranged in the outer ring. There are 12 of them. 2) Small green/grey circles: Count the number of smaller green/grey circles in the inner ring. There are also 12 of these. 3) Orange shapes: Count the number of orange shapes in the very center. There are 12 of these as well. Since all three types of repeating elements (large green circles, small green/grey circles, and orange shapes) appear 12 times and are arranged symmetrically around the center, the pie exhibits 12-fold rotational symmetry. If each child receives an identical piece of pie, then the pie must be divided into sectors that are identical in shape and contain identical patterns. The smallest such sector would be one that contains one of each of the repeating elements. Imagine drawing radial lines from the center of the pie that equally divide the circle. If we divide the pie into 12 equal sectors (each $360/12 = 30$ degrees), each sector will contain: One large green circle. One small green/grey circle. One orange shape. Since these 12 pieces would be identical and cover the entire pie, the maximum possible number of children is 12. If there were fewer children (e.g., 6), each piece would be larger and contain multiple sets of these elements (e.g., two large green circles, two small green circles, two orange shapes), but the question asks for the *maximum* number, which corresponds to the smallest possible identical pieces. Therefore, the maximum number of children is 12. The final answer is **12**.

Ground Truth	Visual Grounding	B > A
	Logical Coherence and Consistency	A > B
	Factual Correctness / No Hallucination	A > B
	Reflection and Exploration	B > A
	Conciseness and Efficiency	A > B

Table 13. Example of an evaluation instance for math reasoning. Both responses appeal to rotational symmetry but arrive at incorrect final answers. Response A is more concise and avoid hallucinated colors. Response B demonstrates slightly better grounding for the inner orange shapes and actively explores alternative solutions, yet contains some logical shifts that reduce coherence.

Verifiable Reasoning Example 2



Question Is the number of green bus greater than the number of blue school bus? Options: (A) Yes (B) No.

Answer A

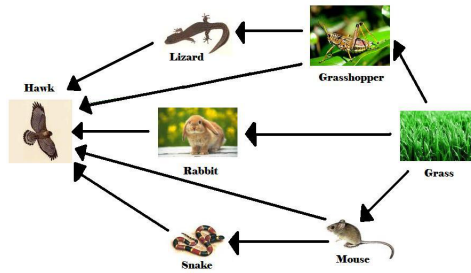
Response A (Gemini-2.5-Flash) 1. Identify green buses: In the image, there is one long green vehicle that is a bus. So, the number of green buses is 1. 2. Identify blue school buses: There are no blue vehicles visible in the image, and certainly no blue school buses. So, the number of blue school buses is 0. 3. Compare the numbers: 1 (green buses) is greater than 0 (blue school buses). Therefore, the number of green buses is greater than the number of blue school buses. The final answer is A

Response B (Gemini-2.5-Flash) Here's a breakdown of the vehicles in the image: - There is one green bus in the **lower right** portion of the image. - There is a **gray dirt bike in the middle-left**. - There is a **golden motorcycle** (possibly a chopper or custom bike) with red and green accents in the upper-middle. - There is a **gray airplane with blue and green accents** in the upper-right. Based on this observation: - Number of green buses = 1 - Number of blue school buses = 0 (**There are no blue buses, nor any vehicles that distinctly look like a school bus of any color, let alone blue.**) Since 1 is greater than 0, the number of green buses is greater than the number of blue school buses. The final answer is A

Ground Truth	Visual Grounding	B > A
	Factual Correctness / No Hallucination	A > B
	Reflection and Exploration	B > A
	Conciseness and Efficiency	A > B

Table 14. Example of an evaluation instance for object counting. Both Gemini-2.5-Flash responses correctly identify more green buses than blue ones. Response A is more **compact**. Response B **provides finer object details and spatial relationships** and **explicitly checks each visible object**, but introduces a minor **hallucination by stating that no school bus is present**, even though one appears in the image.

Verifiable Reasoning Example 3



Question In the diagram of the food web shown, a pointing arrow from A to B indicates that A is the source of energy that helps B survive. If the number of lizard and grasshopper decrease, the population of grass will likely? (A) decreases (B) increases (C) remains the same (D) cannot be determined from the image.

Answer B

Response A (MiMo-VL-7B-RL) <think> I need to analyze this food web diagram to understand the relationships between organisms and determine what would happen to the grass population if lizards and grasshoppers decrease. **Looking at the diagram:** - Grass is eaten by: grasshopper, rabbit, and mouse - Grasshopper is eaten by: lizard and hawk - Lizard is eaten by: hawk - Rabbit is eaten by: hawk - Mouse is eaten by: hawk and snake - Snake is eaten by: hawk If the number of lizards and grasshoppers decreases: 1. Grasshoppers eat grass, so fewer grasshoppers would mean less consumption of grass. 2. Lizards don't directly eat grass, but they eat grasshoppers. With fewer lizards, there would be fewer predators controlling the grasshopper population, which could potentially lead to more grasshoppers in the long run. However, the question states that both lizard and grasshopper populations are decreasing. Since both grasshoppers (which eat grass) and lizards (which control grasshoppers) are decreasing, we have two opposing effects: - Fewer grasshoppers means less grass consumed - Fewer lizards means less predation on grasshoppers, which could lead to more grasshoppers in the future. However, since the question specifically asks about the immediate effect of these decreases, the most direct impact would be that with fewer grasshoppers consuming grass, the grass population would likely increase. </think> In this ecosystem: - Grass is consumed by three primary herbivores: grasshoppers, rabbits, and mice - Grasshoppers are preyed upon by both lizards and hawks - Lizards are preyed upon by hawks. When the number of lizards and grasshoppers decreases: 1. Grasshoppers directly consume grass. With fewer grasshoppers, there would be less consumption of grass. 2. Lizards don't directly consume grass, but they do control the grasshopper population through predation. With fewer lizards, there would be less predation pressure on grasshoppers, which could potentially lead to more grasshoppers in the future. However, the question specifies that both lizard and grasshopper populations are decreasing. The most immediate effect would be reduced grass consumption due to fewer grasshoppers. Since grass is also consumed by rabbits and mice (whose populations aren't mentioned as changing), the reduction in grasshoppers would likely result in increased grass availability. Therefore, the population of grass will likely **(B) increases**.

Response B (Gemini-2.5-Flash) To determine the likely change in the grass population, we need to look at what organisms consume grass in this food web. 1. Identify consumers of Grass: The diagram shows arrows from Grass to Grasshopper, Rabbit, and Mouse. This means Grasshoppers, Rabbits, and Mice all eat Grass. 2. Analyze the given scenario: "the number of lizard and grasshopper decrease". 3. Focus on the direct impact on Grass: - The Grasshopper population decreases. Since grasshoppers eat grass, a decrease in their numbers means there will be fewer organisms consuming grass. - The Lizard population decreases. Lizards eat grasshoppers. **While a decrease in lizards might lead to an increase in grasshoppers (their prey) in some scenarios, the problem explicitly states that grasshoppers also decrease. Therefore, the decrease in lizards doesn't counteract the effect of fewer grasshoppers eating grass.** In fact, fewer lizards might be a consequence of fewer grasshoppers. 4. Conclusion: With fewer grasshoppers eating grass, the predation pressure on the grass population is reduced. Assuming other factors (like Rabbit and Mouse populations, or environmental conditions) remain constant, the grass will be consumed less. This reduced consumption will likely lead to an increase in the grass population. Therefore, if the number of lizard and grasshopper decrease, the population of grass will likely increase. The final answer is **B**.

Ground Truth	Visual Grounding	A > B
	Reflection and Exploration	B > A
	Conciseness and Efficiency	B > A

Table 15. Example of an evaluation instance in judging biological reasoning. Both responses reach the correct conclusion (grass increases). Response A provides a more complete and detailed **description of the food web**. Response B is more **efficient** and together shows **deeper reflection** by explicitly recognizing that the decrease in lizards **does not counteract the effect** of fewer grasshoppers eating grass.

E. Additional Experimental Results

E.1. To Think or Not?

To study how thinking influences judge model behaviors in following diverse criteria, we examine two strong open-source LMM families that are known for producing high-quality reasoning traces across diverse domains. InternVL3.5 [43] supports both thinking (default) and non-thinking modes controlled via system prompts, enabling a clean mode-switch comparison within the same model weights. Qwen3-VL [55] provides two model variants—Instruct and Thinking—the latter specifically developed to enhance reasoning capabilities.

Table 16 presents the results, showing two trends:

1. *Smaller models benefit more from thinking, while larger models show limited gains.*

For the smaller 8B models, both the mode-switch (InternVL3.5) and model-variant (Qwen3-VL) show clear and consistent improvements from enabling thinking, with gains across all metrics on both open-ended and reasoning splits. In contrast, the larger \sim 30B models show more modest effects: thinking brings no significant improvement for InternVL3.5, and for Qwen3-VL it enhances reasoning judgments but slightly reduces performance on open-ended splits.

These results suggest that explicit thinking during judgment helps smaller models better adhere to fine-grained evaluation criteria and produce criterion-specific critic reasoning. For larger models that already internalize such criteria-following capacities and reasoning patterns, additional thinking offers little benefit and may even amplify judge model biases.

2. *Thinking benefits reasoning judgments more.*

For both Qwen3-VL and InternVL families, the thinking judge yields larger gains in evaluating verifiable reasoning than in open-ended generation tasks. Specifically, *Qwen3-VL-Thinking* outperforms its *Instruct* variant at both 8B and 32B scales, demonstrating improved ability to recognize criterion trade-offs and capture preference conflicts. This finding differs from the observation in Sec. 4.2, where RL-finetuned reasoning models on domain-specific data based on Qwen2.5-VL showed no improvement in reasoning judgments and even weakened trade-off recognition.

This indicates that the capacity for reasoning judgment arises from general thinking abilities developed through broad reasoning training across diverse domains, rather than from narrow domain-specific fine-tuning, which often leads to overfitting and reduced generalization.

E.2. Joint Multi-Criterion Judgment

Task Formulation. In our standard setting, each evaluation instance (prompt + response pair) with K applicable criteria requires the LMM judge to perform K separate inferences, each instantiated with a single-criterion evaluation prompt as shown in Table 9. Here, we investigate an alternative setting in which the judge performs *joint multi-criterion judgment*. For each instance, all K criteria are presented to the model simultaneously, and a *single* inference is used to produce criterion-level judgments for all criteria. Table 17 shows the corresponding joint evaluation prompt: the model is instructed to treat each criterion separately while outputting all criterion-level judgments in one pass. All inference hyper-parameters remain identical to the standard single-criterion setup, allowing us to isolate the effect of joint prompting on judge model behavior.

Results and Analysis. We evaluate four top-performing proprietary LMMs—GPT-4o, GPT-5, o4-mini, and Gemini-2.5-Pro—under the joint multi-criterion judgment setting. The results in Table 18 reveal two major observations:

1. *Joint multi-criterion judgment affects models unevenly.*
GPT-4o exhibits a clear performance drop across all metrics, suggesting that combining multiple criteria in a single pass amplifies its internal biases and weakens its ability to follow diverse criteria. GPT-5, in contrast, benefits from the joint setting—showing improved pluralistic accuracy and better alignment with human preferences on conflict cases. For o4-mini and Gemini-2.5-Pro, the effects of joint criterion judgment are mixed on open-ended tasks but consistently drop on reasoning domains.
2. *Joint judgment generally reduces the model’s sensitivity to criterion-level trade-offs.*
Across all models—except GPT-5 in the reasoning split—the trade-off sensitivity decreases. This trend is expected: given the autoregressive nature of LLMs/LMMs, generating multiple criterion-level judgments within a single inference pass inevitably induces inter-criterion dependencies, making the judge model more likely to assign the same preference direction across criteria for the same pair of responses. Table 19 shows a case where *o4-mini* fails to capture criterion-level conflicts under joint criterion judgment.

Model	Open-ended				Verifiable Reasoning			
	Pluralistic	Conflict	Tradeoff	Crit.-Avg.	Pluralistic	Conflict	Tradeoff	Crit.-Avg.
InternVL3.5-8B (no-think)	23.41	24.95	44.66	59.09	27.78	28.83	54.13	61.96
InternVL3.5-8B (think)	25.08	32.34	62.14	61.04	32.54	39.15	69.72	65.85
Δ	+1.67	+7.39	+17.48	+1.95	+4.76	+10.32	+15.59	+3.89
Qwen3-VL-8B-Instruct	18.39	18.36	34.47	54.16	16.67	19.22	35.78	56.92
Qwen3-VL-8B-Thinking	24.75	37.33	66.50	60.61	38.89	51.25	83.49	70.70
Δ	+6.36	+18.97	+32.03	+6.45	+22.22	+32.03	+47.71	+13.78
InternVL3.5-38B (no-think)	29.43	35.93	61.65	65.36	40.48	47.69	75.23	71.37
InternVL3.5-38B (think)	30.43	33.73	64.08	65.10	37.30	47.69	75.23	69.82
Δ	+1.00	-2.20	+2.43	-0.26	-3.18	0.00	0.00	-1.55
Qwen3-VL-32B-Instruct	30.43	40.32	68.93	65.49	39.68	48.75	70.64	71.42
Qwen3-VL-32B-Thinking	29.10	40.12	67.96	64.88	43.65	53.38	80.73	73.88
Δ	-1.33	-0.20	-0.97	-0.61	+3.97	+4.63	+10.09	+2.46

Table 16. Comparison of thinking vs. non-thinking LMM judges. Relative improvements are shown in green and decreases in red.

You are an expert in evaluating the quality of AI-generated responses according to multiple evaluation criteria. Your task is to assess two responses generated by different AI assistants in reply to a user’s question about an image. The image is provided as part of the input.

You must evaluate the responses based on the following $\{K\}$ evaluation criteria. **Analyze each criterion independently and exclusively**—your judgment on one criterion should not influence your judgment on another. Do not consider any other dimensions or criteria beyond what is specified below.

{Criteria}

Here are the inputs for your evaluation:

[Question]: {Question}

[Response 1]: {Response1}

[Response 2]: {Response2}

Instructions: For *each* criterion listed above, you must: (1) provide a detailed justification for your evaluation, referring to specific elements in the responses, how they align with that criterion, and relevant visual details from the image; and (2) on the final line of each criterion’s evaluation, provide your judgment based solely on that criterion. You must choose one response as better; do not indicate a tie.

Format your response as follows for each criterion:

Criterion: [Criterion Name 1]

[Your detailed justification here]

Judgment: Response X is better.

Criterion: [Criterion Name 2]

[Your detailed justification here]

Judgment: Response X is better.

... (continue for all criteria)

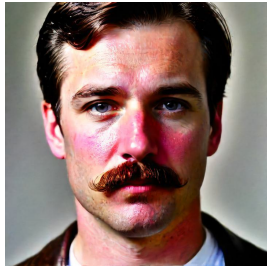
Strictly follow this format: ``Response X is better.`` on the last line of each criterion block.

Table 17. Joint multi-criterion evaluation prompt for LMM judges. All applicable criteria for an evaluation instance are assessed in a single pass, and the judge is explicitly instructed to treat each criterion separately.

Model	Open-ended				Verifiable Reasoning			
	Pluralistic	Conflict	Tradeoff	Criterion Avg.	Pluralistic	Conflict	Tradeoff	Criterion Avg.
GPT-4o	31.44	44.91	66.02	69.57	41.27	55.16	84.40	69.79
+ Joint multi-criterion	30.10	26.35	38.83	64.91	34.13	34.52	58.72	63.15
Δ	-1.34	-18.56	-27.19	-4.66	-7.14	-20.64	-25.68	-6.64
GPT-5	29.77	38.52	62.62	68.51	45.24	56.58	78.90	77.41
+ Joint multi-criterion	34.78	43.51	60.68	70.43	47.62	61.57	82.57	79.58
Δ	+5.01	+4.99	-1.94	+1.92	+2.38	+4.99	+3.67	+2.17
o4-mini	32.78	43.11	64.56	69.67	53.17	65.84	83.49	80.85
+ Joint multi-criterion	36.79	40.12	58.74	68.82	45.24	55.87	77.06	75.61
Δ	+4.01	-2.99	-5.82	-0.85	-7.93	-9.97	-6.43	-5.24
Gemini-2.5-Pro	28.76	37.92	66.50	63.67	41.27	52.33	75.93	73.06
+ Joint multi-criterion	32.11	41.12	62.62	67.06	35.71	40.21	56.88	70.96
Δ	+3.35	+3.20	-3.88	+3.39	-5.56	-12.12	-19.05	-2.10

Table 18. Comparison of joint multi-criterion judgment against standard single-criterion judgment. Relative improvements are shown in green, decreases in red.

Case Study: Joint Multi-Criterion Failure of the *o4-mini* Judge in Capturing Preference Conflicts



Question	Provide a single 10-year age bracket (e.g., “30s,” “40s,” etc.) and describe the man’s hair color and hairstyle based only on the image.	
Response 1	Based on the image provided, the man appears to be in his 30s. His hair is dark brown and is styled in a classic fashion, swept to the side, which complements his period attire. The style is reminiscent of early 20th-century fashion, which aligns with the vintage feel of the image.	
Response 2	The man in the photo appears to be in his late 30s to early 40s. His hair is dark brown with a slight wave pattern. It is neatly styled, suggesting that he takes care of his appearance. The hair is cut short, and the parting on the side complements his overall look.	
Ground Truth	Completeness and Coverage	Resp. 1 > Resp. 2
	Factuality / No Hallucination	Resp. 2 > Resp. 1
Single-Criterion Judgment (<i>o4-mini</i>)	<p>Completeness and Coverage The prompt asks for a single 10-year age bracket (e.g., “30s,” “40s,” etc.). Response 1 clearly states “in his 30s,” satisfying that requirement exactly. Response 2 instead gives “late 30s to early 40s,” which spans two decades and thus does not fulfill the prompt’s decade-long range specification. - Both responses describe the hair color (“dark brown”) and hairstyle. However, Response 1 stays focused on the core question by succinctly noting the side-swept, classic style evident in the image. Response 2 adds extra inferences (“suggesting that he takes care of his appearance”) and splits the age range improperly, making it less precise with respect to the user’s request. Because Response 1 fully and directly addresses both parts of the question—providing the correct decade-long age bracket plus hair color and style—while Response 2 does not, Response 1 is the better answer under the completeness and coverage criterion. Response 1 is better. ✓</p> <p>Factuality / No Hallucination Response 1 introduces unsupported details about “period attire” and “early 20th-century fashion,” neither of which are evident in the image—this is a visual hallucination. Response 2, by contrast, sticks to what is visible (age estimate spanning late 30s to early 40s, dark brown hair, slight wave, side part, neatly styled) without inventing extraneous context. Response 2 is better. ✓</p>	
Joint Multi-Criterion Judgment (<i>o4-mini</i>)	<p>Criterion: Completeness and Coverage Response 1 directly answers both parts of the prompt within the requested format: it gives a single 10-year range (“30s”) and describes hair color (“dark brown”) and style (“classic fashion, swept to the side”), even tying in contextual cues (“period attire,” “vintage feel”). Response 2, by contrast, straddles two decades (“late 30s to early 40s”), which fails to provide a single 10-year band as requested, and although it describes hair color and neat styling, it does not anchor its observations as neatly to the prompt’s format. Judgment: Response 1 is better. ✓</p> <p>Criterion: Factuality / No Hallucination Both responses correctly observe dark brown hair and a side part, but Response 2 introduces imprecision by spanning “late 30s to early 40s,” which contradicts the user’s instruction for a single decade. Response 1 sticks to a plausible age (“30s”) based on the man’s appearance. Neither response hallucinates details about color, but Response 2’s age estimate violates the prompt’s constraints. Judgment: Response 1 is better. ✗</p>	

Table 19. In this case, *o4-mini* correctly identifies the completeness vs. no-hallucination conflict under single-criterion judgments but collapses to a uniform preference under joint multi-criterion judgment. With single-criterion evaluation prompts, the model successfully adheres to each criterion independently and captures the preference disagreement, whereas under joint prompting, completeness dominates and the model fails to follow the no-hallucination criterion.