CVPR JUNE 3-7, 2026 DENVER COLORADO

Multi-Crit: Benchmarking Multimodal Judges on Pluralistic Criteria-Following

CVPR 2026 (Highlight)

Tianyi Xiong*, Ellen Yi-Ge*, Ming Li, Zuolong Zhang, Pranav Kulkarni, Kaishen Wang, Qi He, Zeying Zhu, Chenxi Liu, Ruibo Chen, Tong Zheng, Yanshuo Chen, Xiyao Wang, Renrui Zhang, Wenhu Chen, Heng Huang

Overview

- Background and Motivation
- Benchmark Design and Data Annotation
- Experiments and Analysis

Why LMM-as-a-Judge? Automated Evaluation

Art & Design

Question: Among the following harmonic intervals, which one is constructed incorrectly?

Options:

(A) Major third <*image 1*>

(B) Diminished fifth <*image 2*>

(C) Minor seventh <*image 3*>

(D) Diminished sixth <*image 4*>

Subject: Music; **Subfield:** Music;
Image Type: Sheet Music;
Difficulty: Medium

Question: Imagine the fragrance of the fruits in the image. How would you describe this to someone who has never had this fruit before?

Response: The fragrance of the mangosteens in the image can be described as sweet and slightly floral, with a hint of citrus aroma. It is a delicate and pleasant smell that entices you to try the fruit.

Multiple-choice & Short-Answer QAs:
verifiable ✅

Open-ended Questions In-the-Wild:
more challenging 😊, hard to verify 💔

LMM-as-a-Judge Paradigm

Input: image, question, response (s), Evaluation prompt,
Output: Reason, Score

Multimodal Input for LMM:
Question: What are the specifics visible in the image?

Setting 1: Pointwise Scoring
Response: The image shows a small train with four red cars, traveling on a track. The train is located in a park setting, and there are potted plants nearby.
Evaluation Prompt: From 0 to 100, how much do you rate for this Text Caption in terms of the correct and comprehensive description of the image? Do not dominant the rating by a single attribute such as recognition correctness, but a overall rating on the object/scene appearance, position, pose, action, shape, etc., and contents in the background. Do not consider the appropriateness or sensitive descriptors, such as "middle-aged western man", judge based on if it has correct specifications of the object and scenes in image. Provide a few lines for explanation and the rate number at last after "Final Score:".
LLaVA-Critic: The text caption provides a generally accurate description of the image. It correctly identifies the small train with four red cars and mentions the park setting and the presence of potted plants. However, there are a few details that could be improved for a more comprehensive description. For instance, the caption could mention that the train is stationary rather than traveling, and it could describe the surrounding structures, such as the building and the wooden fence, to give a fuller picture of the scene. Final Score: 85.

Xiong, Tianyi, et al. "Llava-critic: Learning to evaluate multimodal models." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.

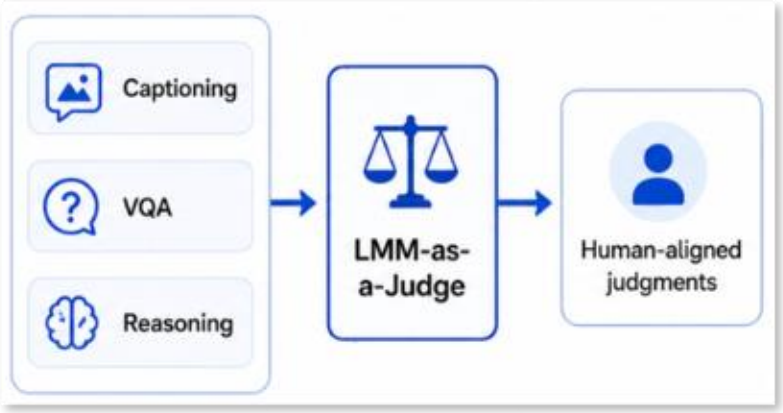
Why is LMM-as-a-Judge popular?

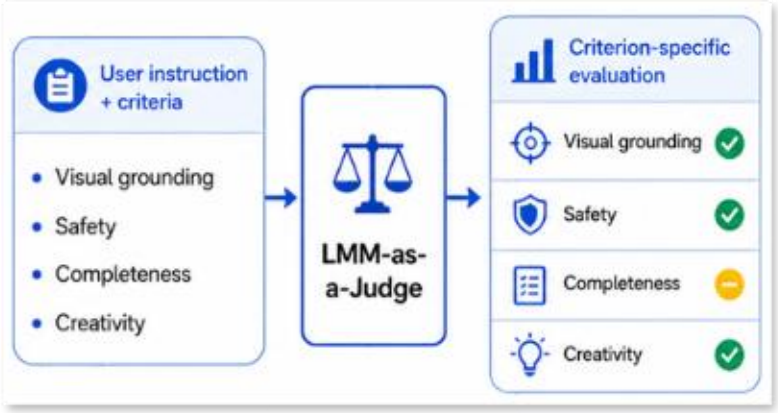
Reliable judgment at scale

Provide human-aligned evaluation across diverse multimodal tasks.

Flexible evaluation interface

Follow user-specified instructions and evaluation criteria.





LMM-as-a-Judge is flexible

MMHal-Bench: Hallucination Detection

Help me evaluate whether the response by the LMM is informative, and whether hallucination exists in it, based on the comparison between the LMM's response and the factual information provided in the image contents, question, and the standard human-generated answer below.

To evaluate the LMM responses, first, begin your evaluation by providing a short explanation. Second, after providing your explanation, you must rate the response by choosing from the following options:

- Rating: 6, very informative with good analysis or reasoning, no hallucination
- Rating: 5, very informative, no hallucination
- Rating: 4, somewhat informative, no hallucination
- Rating: 3, not informative, no hallucination
- Rating: 2, very informative, with hallucination
- Rating: 1, somewhat informative, with hallucination
- Rating: 0, not informative, with hallucination

DriveBench: Autonomous Driving

Please evaluate the predicted answer on a scale from 0 to 100, where a higher score reflects precise alignment with the correct answer and well-supported reasoning. Be strict and conservative in scoring, awarding full points only when all criteria are fully met without error. Deduct points for minor inaccuracies, omissions, or lack of clarity. Distribute the Total Score across the following criteria:

1. Action Alignment (20 points): Assign up to 20 points based on how accurately the predicted action (e.g., forward, turn left, turn right) matches the correct answer.

2. Motion Precision (20 points): Award up to 20 points based on how closely the predicted motion (e.g., speed up, decelerate) aligns with the correct motion in the answer. motion prediction is provided.

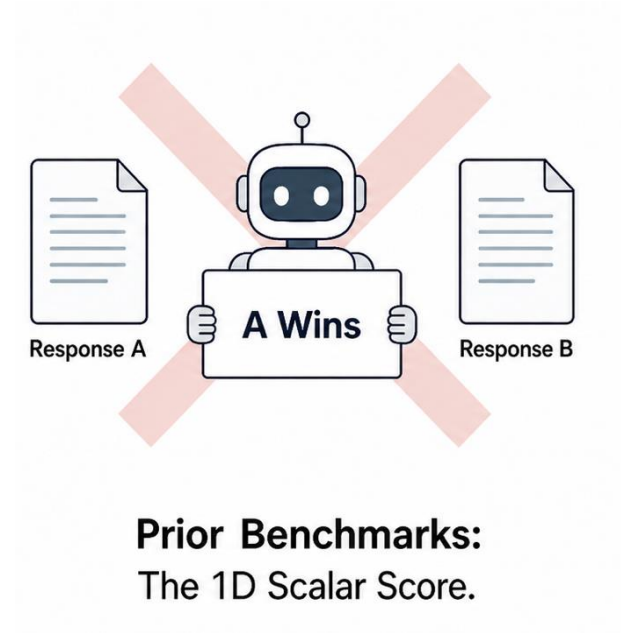
3. Driving Context Appropriateness (15 points): Score up to 15 points for the relevance of the predicted answer to the driving context implied by the correct answer, emphasizing logical alignment with the situation.

4. Situational Awareness (15 points): Award up to 15 points for demonstrated awareness of environmental factors (e.g., traffic participants, obstacles) relevant to the action or motion.

...

Limitation of Existing Benchmarks

Existing reward benchmarks focus on how well judges align with overall human preference, but overlook their ability to follow user-defined evaluation instructions.



VL-RewardBench Leaderboard

https://vl-rewardbench.github.io/

Leaderboard | Data Viewer | About

Dataset Split: test

Sample Index: 93

Total Samples: Total samples: 1247

Sample Image

Sample ID: RLAIF-V-40207

Chosen Response ✅

The main objects or subjects in the image are a group of people, some of whom are playing Frisbee, and a man jumping to catch the Frisbee. This scene takes place in a crowded park.

Rejected Response ❌

The main objects in the image are a man playing Frisbee, a crowd of people watching him, and a blue Frisbee in the air. The image captures the excitement and action of the sport, with the man jumping into the air to catch the Frisbee.

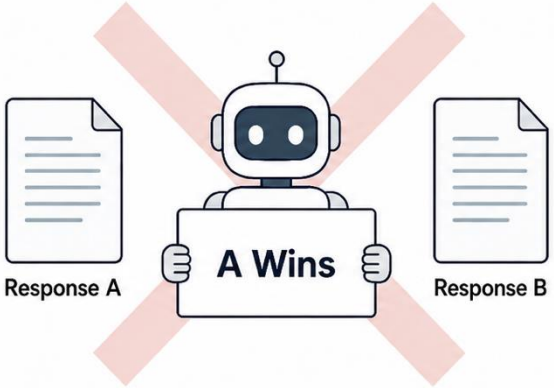
Query

What are the main objects or subjects in the image? Please describe them in detail.

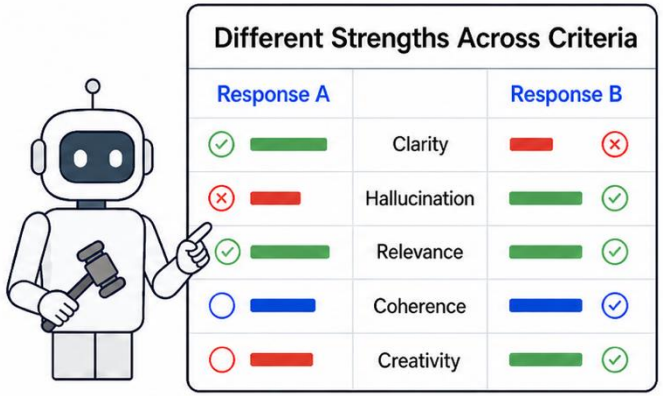
Li, Lei, et al. "VL-RewardBench: a challenging benchmark for vision-language generative reward models." *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025.

Limitation of Existing Benchmarks

Existing reward benchmarks focus on how well judges align with overall human preference, but overlook their ability to follow user-defined evaluation instructions.



Prior Benchmarks:
The 1D Scalar Score.



The Reality:
Multi-dimensional Judgments.

Can LMM-based judges follow diverse evaluation criteria and recognize trade-offs and conflicts among criterion-level judgments?

⚖ Multi-Crit

✅ Challenging response pairs beyond overall preference

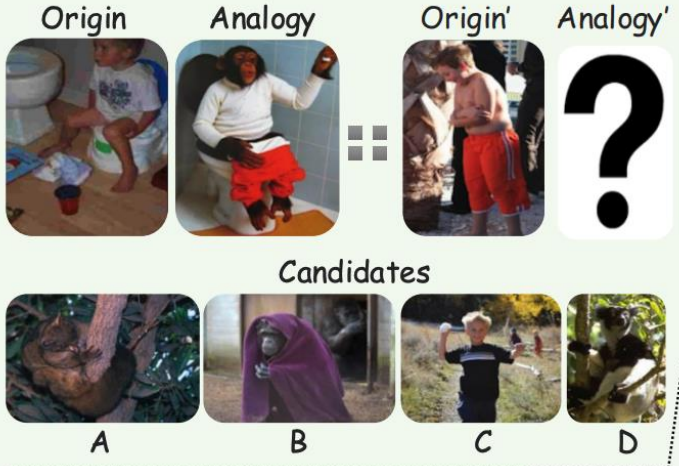
✅ Criterion-level human annotations revealing preference conflicts

Query: Select the correct candidate.

Response A: …relationship between "Origin" and "Analogy" involves a human action being mimicked by a chimpanzee … in "Origin", the child is standing with a towel, possibly drying off after swimming … best candidate that mirrors this in a similar way is B, where the chimpanzee is also wrapped in a towel, resembling the action….

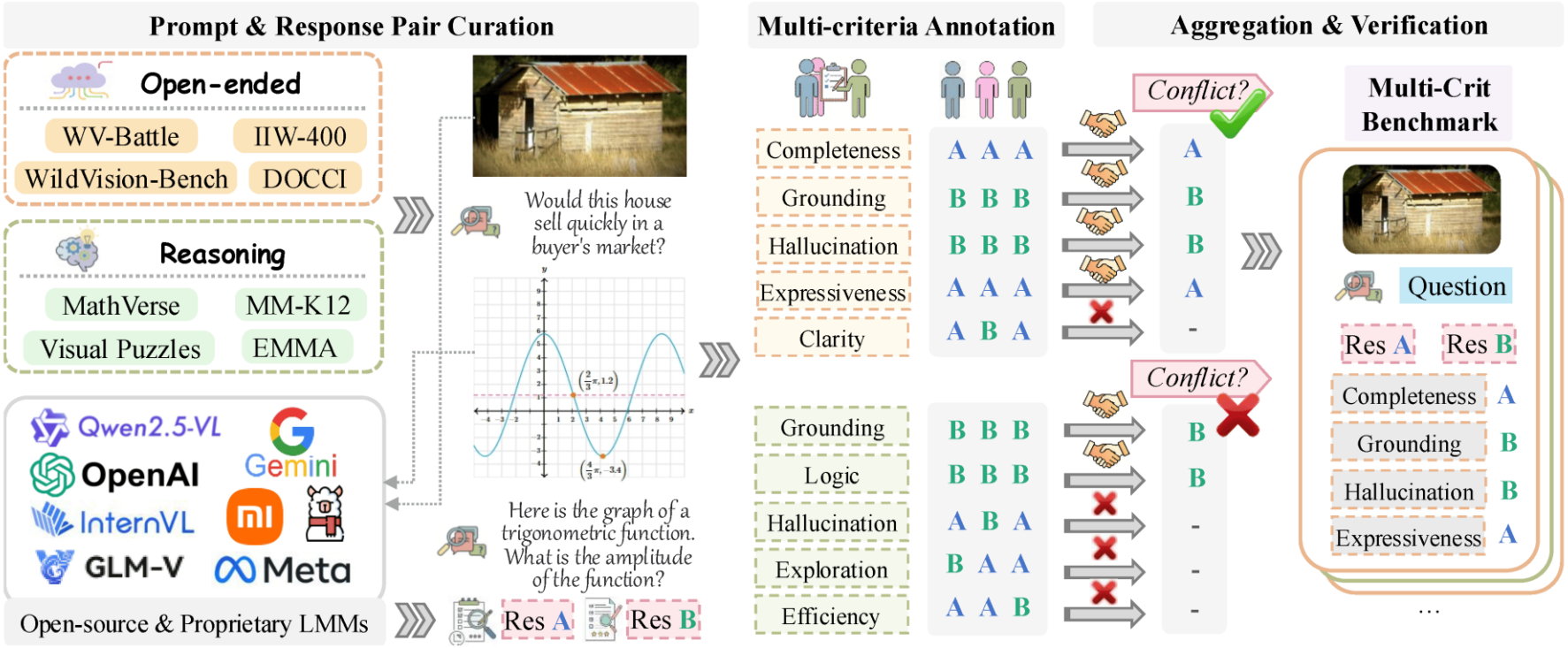
Response B: … child in "Origin" is mimicked by the chimpanzee in "Analogy", but with a different setting (child sitting on a toilet, chimpanzee on a chair) … need to find chimpanzee mimicking the boy standing with his hands on his hips. Check responses: A: koala climbing a tree. B: A chimpanzee wrapped in a blanket., … no candidate exactly matches the boy's pose, however, the closest match in terms of the animal being a chimpanzee is candidate B….

VL-RewardBench, MM-RLHF, MM-RewardBench — A ? B

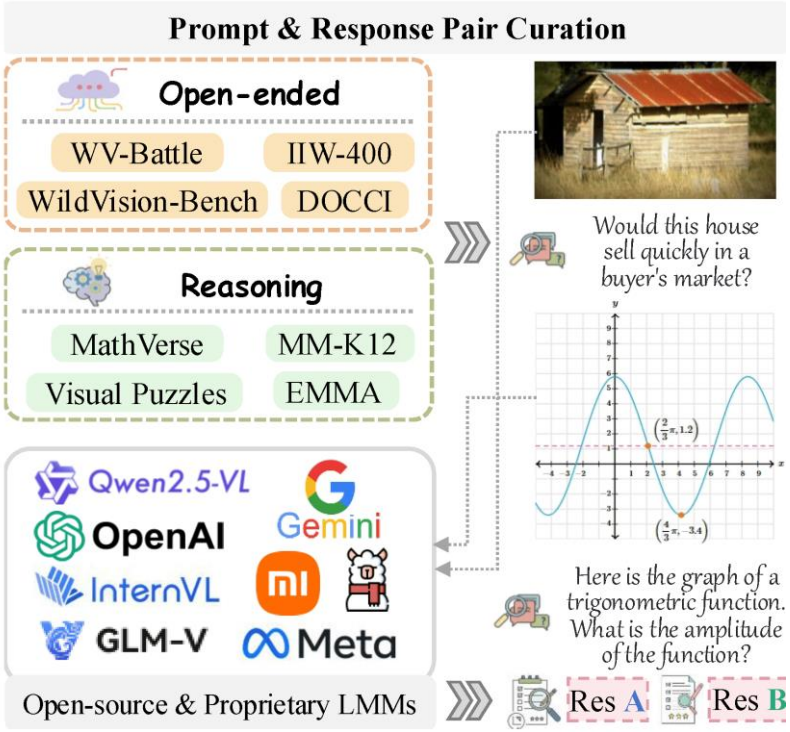


Ours	A	B
Visual Grounding	✗	✓
Logical Coherence and Consistency	✗	✓
Factuality / No Hallucination	✓	✗
Reflection and Exploration	✗	✓
Conciseness and Efficiency	✓	✗

How We Build Multi-Crit



How We Build Multi-Crit



Two Judgment Scenarios

- **Open-ended generation:** where fixed metrics are limited.
- **Verifiable reasoning:** evaluation on model-generated reasoning processes.

Challenging Response pairs

- **Pre-filtering:** select pairs with subtle yet meaningful quality differences.

Constructing Criteria

- Surveyed existing benchmarks that use LMM judges and analyzed their evaluation prompt designs
- Three design principles:
  - Practicality: reflect common use cases of LMM judges
  - Specificity: capture distinct response-quality dimensions
  - Generality: evaluate fundamental model behaviors, not prompt-specific content
- 5 criteria for each evaluation setting

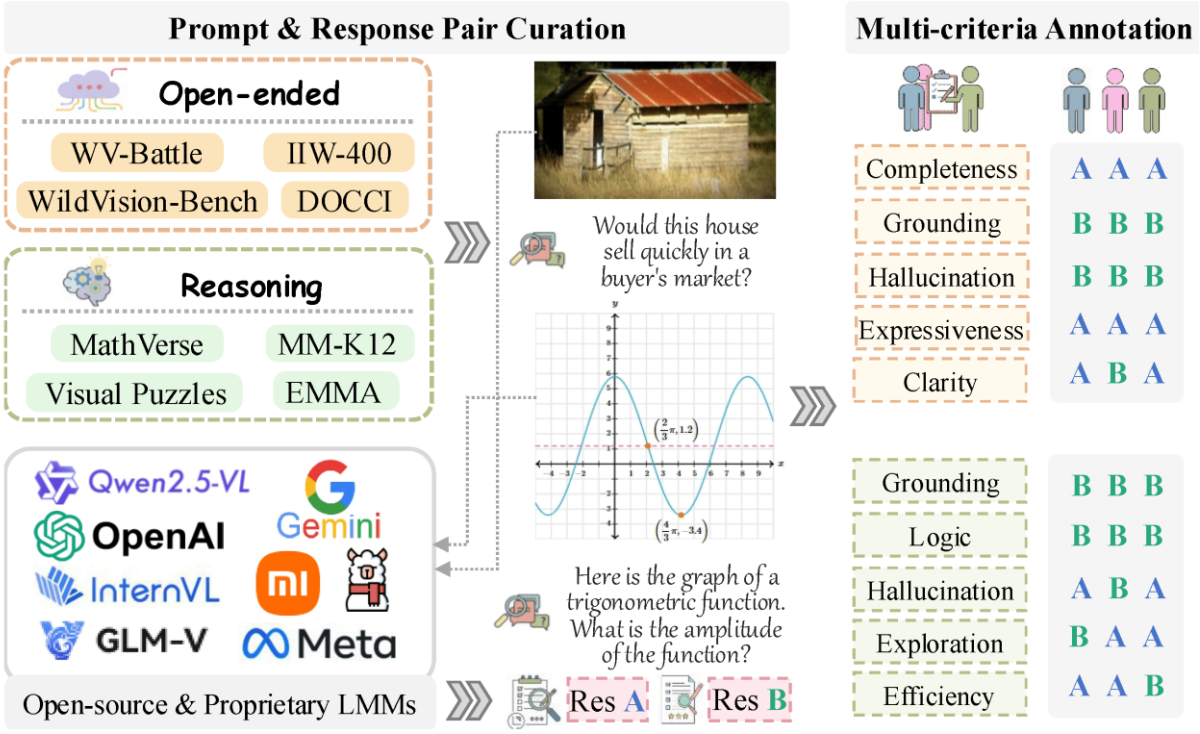
`Criteria for Judging Open-ended Generation`

Completeness and Coverage: Address the full scope of the task in the user's query, covering all major elements specified in the prompt as well as relevant visual aspects and contextual cues.
Visual Grounding and Details: Reference observable elements in the image such as objects, spatial relationships, colors, or text, and bases its description or analysis on these details.
Factuality / No Hallucination: Avoid visual or factual errors, ensuring all details and claims are presented in the image or reasonably supported by the prompt.
Creativity and Expressiveness: Demonstrates imagination and originality when appropriate, or precise and knowledgeable articulation for analytical tasks, while remaining contextually appropriate.
Clarity and Coherence: Communicates ideas clearly and logically, with fluent language, well-organized structure, and smooth flow of information.

`Criteria for Judging Verifiable Reasoning`

Visual Grounding: Reference important visual elements—such as objects, layout, or text—and integrates them meaningfully into the reasoning.
Logic Coherence and Consistency: Follow a clear, step-by-step logic without contradictions or unjustified leaps, and ensure the answer aligns with reasoning.
Factuality / No Hallucination: Ensure accuracy of all claims and support them with the input, avoiding hallucinated visual details or factual errors.
Reflection and Exploration: Demonstrate depth of reasoning through reflection and exploration of alternative interpretations, particularly in complex tasks.
Conciseness and Efficiency: Remains concise and focused, matching the task complexity while avoiding redundancy or unnecessary over-analysis.

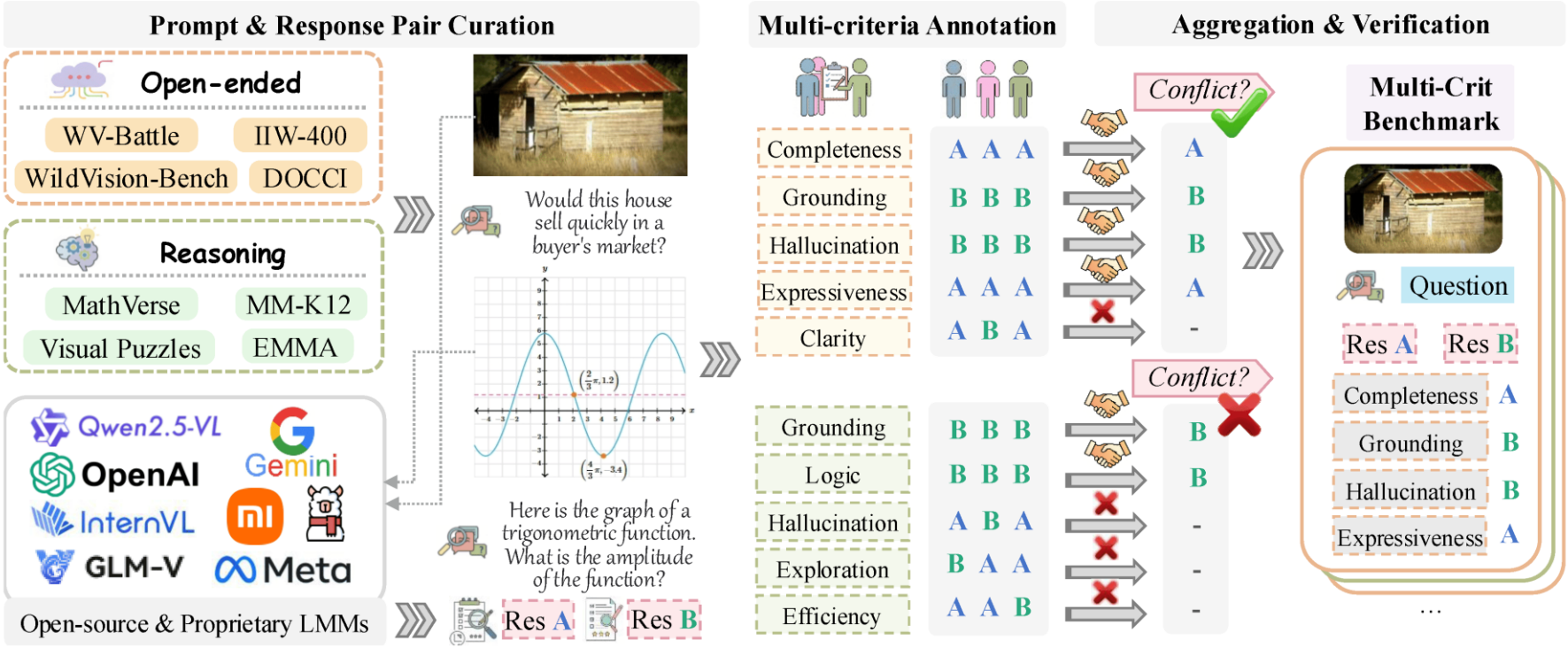
How We Build Multi-Crit



Human Annotation

- Annotators: PhDs in AI
- Each criterion is independently assigned to 3 annotators
- Iterative discussions are conducted to align annotation standards and refine the annotation guidelines

Verification and Aggregation



Data Statistics

Statistic	Number
Open-ended	
- Prompts	299
- with conflicting preferences	206 (68.9%)
- Criterion-level judgment	1,000
- Conflicting criterion pairs	501
Verifiable Reasoning	
- Prompts	126
- with conflicting preferences	109 (86.5%)
- Criterion-level judgment	425
- Conflicting criterion pairs	281
Question length quantiles	(5, 10, 27)
Response length quantiles	(104, 163.5, 247)

Table 1. Multi-Crit's key statistics.

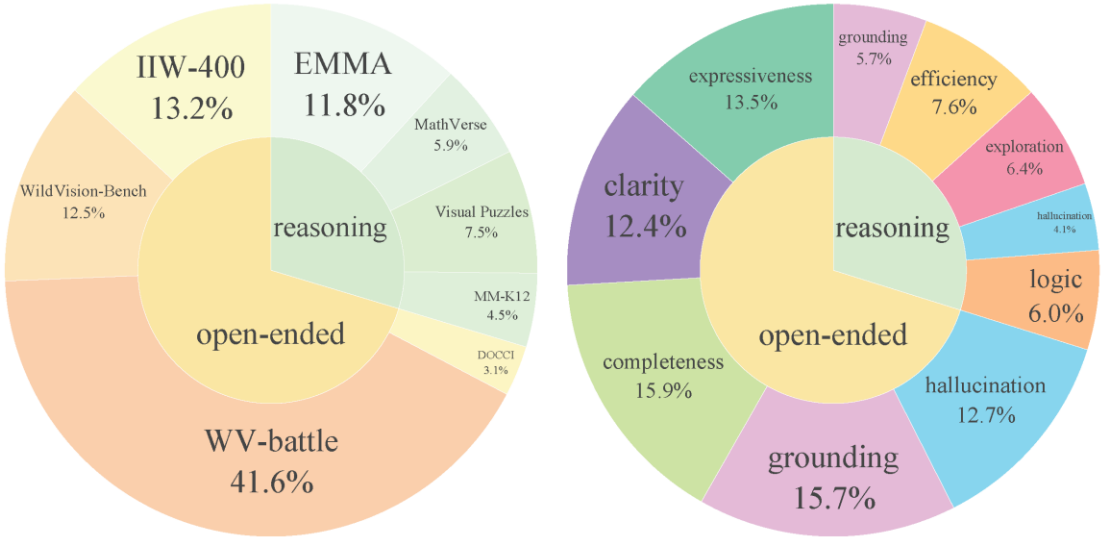


Figure 3. Distribution of prompt sources (left) and evaluation criteria (right).

425 response pairs, 1,425 criterion-level judgments, and 782 conflicting criterion pairs.

Metrics for Evaluating Criteria Following

Pluralistic Accuracy (PAcc)

Gets all criteria correct.

Checks whether the judge gets all criteria correct for each evaluation instance.

Trade-off Sensitivity (TOS)

Detects criterion-level trade-offs.

Assesses whether the judge detects at least one trade-off between the two responses.

Conflict Matching Rate (CMR)

Resolves conflicting criteria correctly.

Measures whether the judge correctly resolves each conflicting criterion pair.

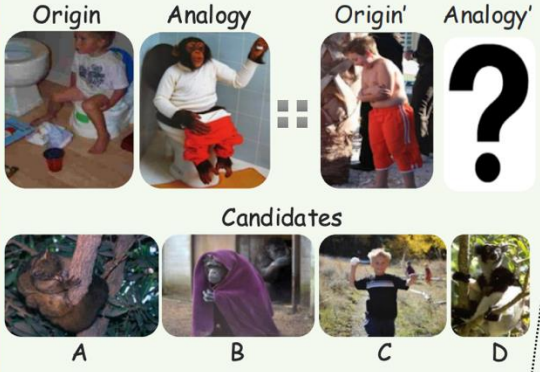
Query: Select the correct candidate.

Response A: …relationship between "Origin" and "Analogy" involves a human action being mimicked by a chimpanzee … in "Origin", the child is standing with a towel, possibly drying off after swimming … best candidate that mirrors this in a similar way is B, where the chimpanzee is also wrapped in a towel, resembling the action….

Response B: … child in "Origin" is mimicked by the chimpanzee in "Analogy", but with a different setting (child sitting on a toilet, chimpanzee on a chair) … need to find chimpanzee mimicking the boy standing with his hands on his hips. Check responses: A: koala climbing a tree. B: A chimpanzee wrapped in a blanket., … no candidate exactly matches the boy's pose, however, the closest match in terms of the animal being a chimpanzee is candidate B….

VL-RewardBench, MM-RLHF, MM-RewardBench

A ? B



Ours	A	B
Visual Grounding	✗	✓
Logical Coherence and Consistency	✗	✓
Factuality / No Hallucination	✓	✗
Reflection and Exploration	✗	✓
Conciseness and Efficiency	✓	✗

	Judge 1	Judge 2	Judge 3	Judge 4
Grounding (B > A)	B > A ✓	B > A ✓	B > A ✓	B > A ✓
Logic (B > A)	B > A ✓	B > A ✓	B > A ✓	B > A ✓
Hallucination (A > B)	B > A ✗	A > B ✓	B > A ✗	A > B ✓
Exploration (B > A)	B > A ✓	A > B ✗	B > A ✓	B > A ✓
Efficiency (A > B)	B > A ✗	B > A ✗	A > B ✓	A > B ✓
Average	0.6	0.6	0.8	1.0
Pluralistic (PAcc ↑)	0.0	0.0	0.0	1.0
Trade-off (TOS ↑)	0.0	1.0	1.0	1.0
Conflict (CMR ↑)	0.0	2/6=0.33	3/6=0.5	1.0

Model	Prompt-Level			Criterion-Level					
	Pluralisic	Conflict	Tradeoff	Completeness	Grounding	Hallucination	Expressiveness	Clarity	Avg
Open-Source LMMs									
Qwen2.5-VL-7B-Instruct	9.41	17.28	36.14	56.12	51.70	48.20	64.12	51.82	54.39
LLaVA-OneVision-7B	11.46	12.77	37.86	53.20	48.80	43.10	55.68	52.26	50.61
Llama-3.2-11B-Vision-Instruct	12.71	15.37	41.26	55.75	44.64	48.07	61.14	46.59	51.24
Molmo-7B	17.06	15.97	34.47	48.23	44.20	55.80	58.03	50.57	51.37
Qwen3-VL-8B-Instruct	18.39	18.36	34.47	56.19	47.77	59.12	55.44	52.27	54.16
MiniCPM-V-4.5-8B	18.73	14.77	27.67	53.10	46.43	53.59	65.80	50.00	53.78
Eagle2.5-8B	20.40	23.75	50.97	64.16	54.91	52.49	59.59	54.55	57.14
GLM-4.1V-9B-Thinking	24.08	30.54	55.34	59.73	55.80	61.33	67.36	51.14	59.07
InternVL3.5-8B	25.08	32.34	62.14	61.50	59.38	56.35			
InternVL3-8B-Instruct	26.09	30.34	61.17	64.60	62.95	56.35			
Qwen2.5-VL-72B-Instruct	28.43	35.53	60.68	69.47	63.39	58.56			
MiMo-VL-7B	29.10	39.52	65.53	65.93	62.95	64.09			
InternVL3-78B-Instruct	29.10	32.53	56.31	73.01	65.62	56.35			
InternVL3.5-38B	30.43	33.73	64.08	71.68	65.18	62.98			
Finetuned Judge Models									
LLaVA-Critic-7B (LlaVA-OV)	11.70	14.17	43.20	56.93	47.32	29.82			
R1-Reward-7B (Qwen2.5-VL)	17.73	20.36	45.63	59.29	60.71	49.72			
UnifiedReward-7B(Qwen2.5-VL)	18.06	8.38	16.50	57.96	52.23	52.49			
LLaVA-Critic-R1-7B(Qwen2.5-VL)	18.39	17.96	39.32	55.31	57.59	46.96			
Proprietary LMMs									
Gemini-2.5-Flash	25.42	31.34	62.14	64.60	60.71	66.30	63.73	57.39	62.55
Gemini-2.5-Pro	28.76	37.92	**66.50**	65.93	59.82	70.17	62.18	60.23	63.67
GPT-5	29.77	38.52	62.62	69.91	69.64	**75.69**	58.55	**68.75**	68.51
o3	31.10	42.71	62.62	66.37	**74.55**	72.93	63.21	**68.75**	69.16
GPT-4o	31.44	**44.91**	66.02	68.14	71.88	65.75	**76.17**	65.91	69.57
Claude-3.7-Sonnet	31.77	42.32	64.08	71.68	**74.55**	63.54	66.84	60.23	67.37
o4-mini	**32.78**	43.11	64.56	**76.11**	**74.55**	65.75	69.43	62.50	**69.67**

Table 3. **Results on open-ended split.** Criterion-level judgment performance is summarized by mean accuracy. The best and second best results are shown in **bold** and underlined respectively. Notably, the top performing *o4-mini* only achieves 32.78% in pluralistic accuracy.

Key Finding 1:

Strongest proprietary models struggle to **maintain consistent pluralistic criteria following.**

Model	Prompt-Level			Criterion-Level					
	Pluralisic	Conflict	Tradeoff	Completeness	Grounding	Hallucination	Expressiveness	Clarity	Avg
Open-Source LMMs									
Qwen2.5-VL-7B-Instruct	9.41	17.28	36.14	56.12	51.70	48.20	64.12	51.82	54.39
LLaVA-OneVision-7B	11.46	12.77	37.86	53.20	48.80	43.10	55.68	52.26	50.61
Llama-3.2-11B-Vision-Instruct	12.71	15.37	41.26	55.75	44.64	48.07	61.14	46.59	51.24
Molmo-7B	17.06	15.97	34.47	48.23	44.20	55.80	58.03	50.57	51.37
Qwen3-VL-8B-Instruct	18.39	18.36	34.47	56.19	47.77	59.12	55.44	52.27	54.16
MiniCPM-V-4.5-8B	18.73	14.77	27.67	53.10	46.43	53.59	65.80	50.00	53.78
Eagle2.5-8B	20.40	23.75	50.97	64.16	54.91	52.49	59.59	54.55	57.14
GLM-4.1V-9B-Thinking	24.08	30.54	55.34	59.73	55.80	61.33	67.36	51.14	59.07
InternVL3.5-8B	25.08	32.34	62.14	61.50	59.38	56.35	69.43	58.52	61.04
InternVL3-8B-Instruct	26.09	30.34	61.17	64.60	62.95	56.35	67.88	56.25	61.61
Qwen2.5-VL-72B-Instruct	28.43	35.53	60.68	69.47	63.39	58.56	70.98	56.82	63.84
MiMo-VL-7B	29.10	39.52	65.53	65.93	62.95	64.09	70.47	53.41	63.37
InternVL3-78B-Instruct	29.10	32.53	56.31	73.01	65.62	56.35	69.43	58.52	61.04
InternVL3.5-38B	30.43	33.73	64.08	71.68	65.18	62.98	63.73	61.93	65.10
Finetuned Judge Models									
LLaVA-Critic-7B (LlaVA-OV)	11.70	14.17	43.20	56.93	47.32	29.82	50.28	44.08	45.69
R1-Reward-7B (Qwen2.5-VL)	17.73	20.36	45.63	59.29	60.71	49.72	55.44	53.98	55.83
UnifiedReward-7B(Qwen2.5-VL)	18.06	8.38	16.50	57.96	52.23	52.40	63.73	55.11	55.74
LLaVA-Critic-R1-7B(Qwen2.5-VL)	18.39	17.96	39.32	55.31	57.59	46.96	63.73	55.11	55.74
Proprietary LMMs									
Gemini-2.5-Flash	25.42	31.34	62.14	64.60	60.71	66.30	63.73	57.39	62.55
Gemini-2.5-Pro	28.76	37.92	**66.50**	65.93	59.82	70.17	62.18	60.23	63.67
GPT-5	29.77	38.52	62.62	69.91	69.64	**75.69**	58.55	**68.75**	68.51
o3	31.10	42.71	62.62	66.37	**74.55**	72.93	63.21	**68.75**	69.16
GPT-4o	31.44	**44.91**	66.02	68.14	71.88	65.75	**76.17**	65.91	69.57
Claude-3.7-Sonnet	31.77	42.32	64.08	71.68	**74.55**	63.54	66.84	60.23	67.37
o4-mini	**32.78**	43.11	64.56	**76.11**	**74.55**	65.75	69.43	62.50	**69.67**

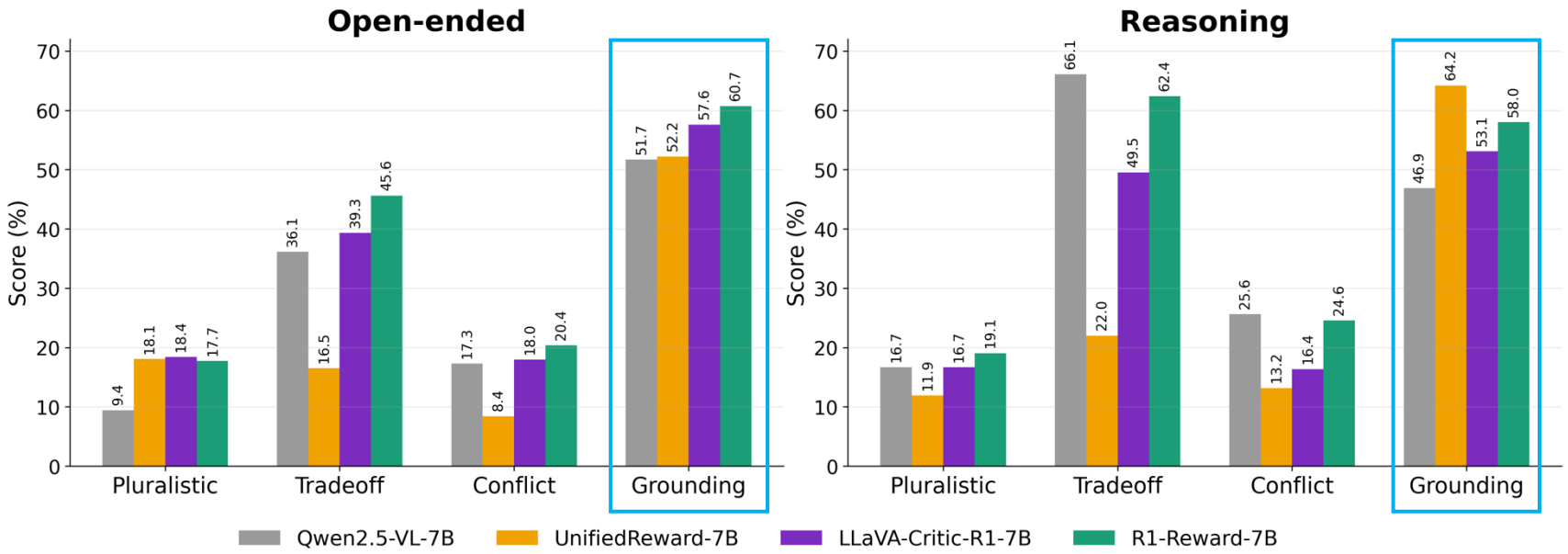
Table 3. **Results on open-ended split.** Criterion-level judgment performance is summarized by mean accuracy. The best and second best results are shown in **bold** and underlined respectively. Notably, the top performing *o4-mini* only achieves 32.78% in pluralistic accuracy.

Key Finding 2:

Open-source LMMs lag behind not only in **criterion-level accuracy** but also in their ability to **recognize trade-offs** and **preference conflicts**.

Key Finding 3:

Critic models fine-tuned on single-preference critic data **improve judgment** consistency in **visual grounding** but generalize **poorly to pluralistic, criterion-driven** evaluation.



What Are the Upper Bounds of Current LMM Judges?

- Proprietary models' **upper-bound** judgment patterns **align closely** with inter-human agreement
- Open-source models show **weaker, less consistent alignment**, reflecting limited capacity to internalize evaluation criteria
  → would **benefit from** scaling high-quality human preference annotations.

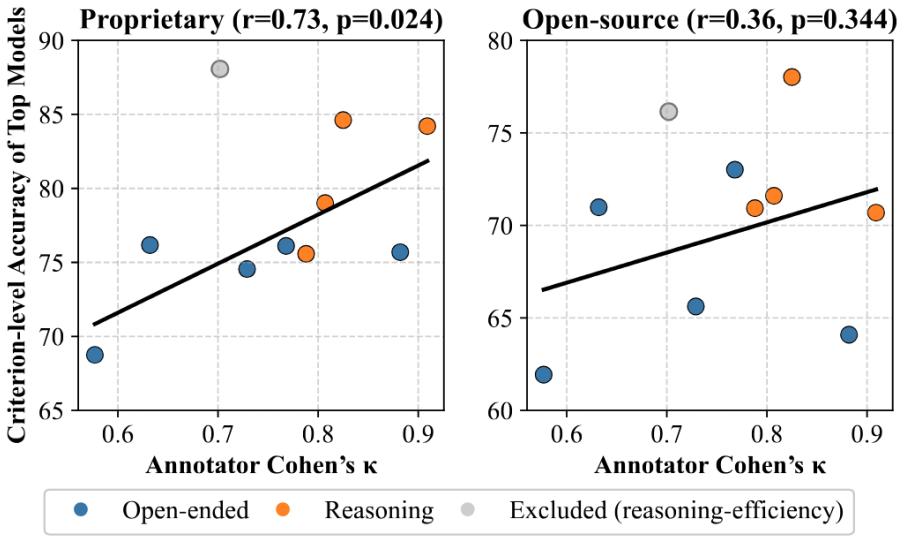
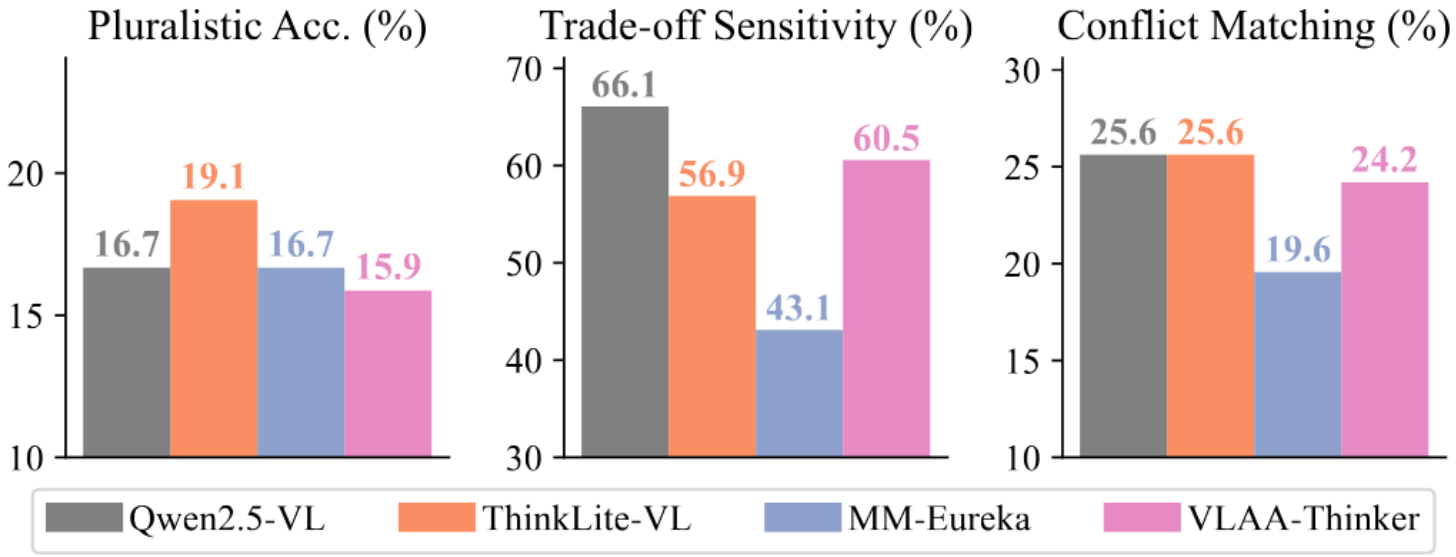


Figure 7. Correlation between criterion-level top judge accuracy and inter-annotator agreement. Each point denotes one criterion.

Analysis – Does reasoning improve judging capacity?



Domain-specific reasoning fine-tuning **does not consistently improve** reasoning judgment and **weakens trade-off recognition**.

Analysis – Does reasoning improve judging capacity?

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

Models with **general thinking capacities** benefit from explicit reasoning during judgment, with larger gains for **smaller models** and **reasoning-judgment** tasks.

Key Takeaways & Industry Relevance

- Pluralistic criteria-following remains challenging for LMM judges.
- The best judge model differs across evaluation criteria. Judge selection and deployment should therefore be criterion-specific.
- Specialized judges and critics trained on holistic preferences can improve visual-grounding evaluation but struggle to generalize to fine-grained criteria-following.

CVPR JUNE 3-7, 2026 DENVER COLORADO

UNIVERSITY OF MARYLAND 18 56

UNIVERSITY OF WATERLOO CONCORDIA CUM VERITATE

Thanks!

Project Page: https://multi-crit.github.io/

Contact: txiong23@umd.edu