ICML
International Conference
On Machine Learning

KAIST CVML

Mitigating Perceptual Judgment Bias in Multimodal LLM-as-a-Judge via Perceptual Perturbation and Reward Modeling

Seojeong Park*, Jiho Choi*, Junyong Kang, Seonho Lee, Jaeyo Shin, Hyunjung Shim†

KAIST AI
KRAFTON

Can Multimodal Judges Really Judge What They See?

- Multimodal LLMs are replacing costly human annotations as **automated evaluators**.
- But judging vision-language responses requires more than fluent reasoning.
- A judge must verify whether the response is visually grounded.

Key Question: Do MLLM judges actually use visual evidence when evaluating responses?

The Core Problem:

Perceptual Judgment Bias

Definition: A multimodal judge fails to penalize a response whose visual claims contradict the image.

a. Insufficient Perceptual Capability

The judge itself misperceives the image,
leading it to reward a visually incorrect answer
simply because it sounds logical.

b. Response-Anchored Judgment Reasoning

The judge can perceive the image correctly,
but still follows the response text during judgment,
ignoring its own accurate perception.

Key Distinction from Prior Work: Not just "poor visual perception" or "hallucination."
This is a *judgment-specific bias* where the judge fails to connect perception with evaluation.

Bias Decomposition:
The Main Error Is Not Just Poor Perception

Decomposing perceptual judgment errors reveals that
anchoring on textual responses is a more significant issue than failing to see the image correctly.

Model	Acc ↑	Error ↓		
		Mode (a) Insufficient Perception	Mode (b) Response-Anchoring	Overall
Qwen2.5-VL-7B-Instruct	69.5%	14.0%	16.4%	30.5%
Flex-Judge-VL-7B	76.6%	9.4%	14.1%	23.5%
Perception-Judge-Flex-7B (Ours)	85.7%	6.7%	7.6%	14.3%

Main message: Response anchoring (14.1%) accounts for a larger portion of errors
than pure perception failure (9.4%) in baseline models.

Controlled Perturbation Analysis: Fluent Reasoning Can Hide Visual Errors

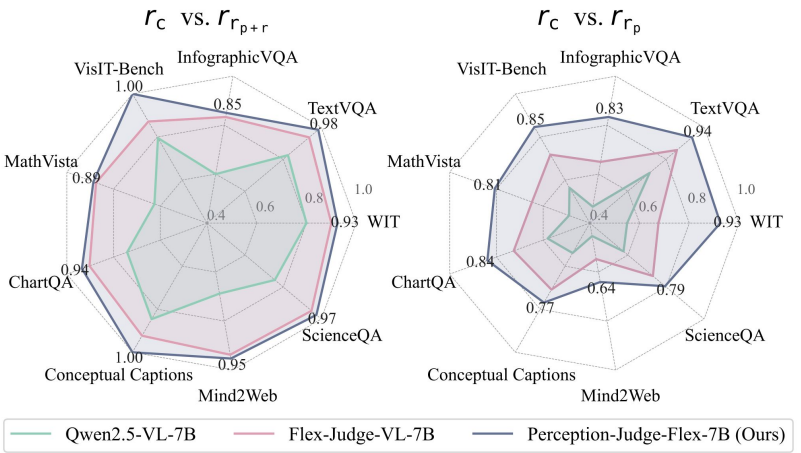
Comparing correct responses (r_c) with perturbed negative responses reveals a key vulnerability in current judges.

[Left] High Accuracy (easy case)

r_c vs. r_{p+r} — Contains visual error + reasoning error.

[Right] Accuracy Drops (hard case)

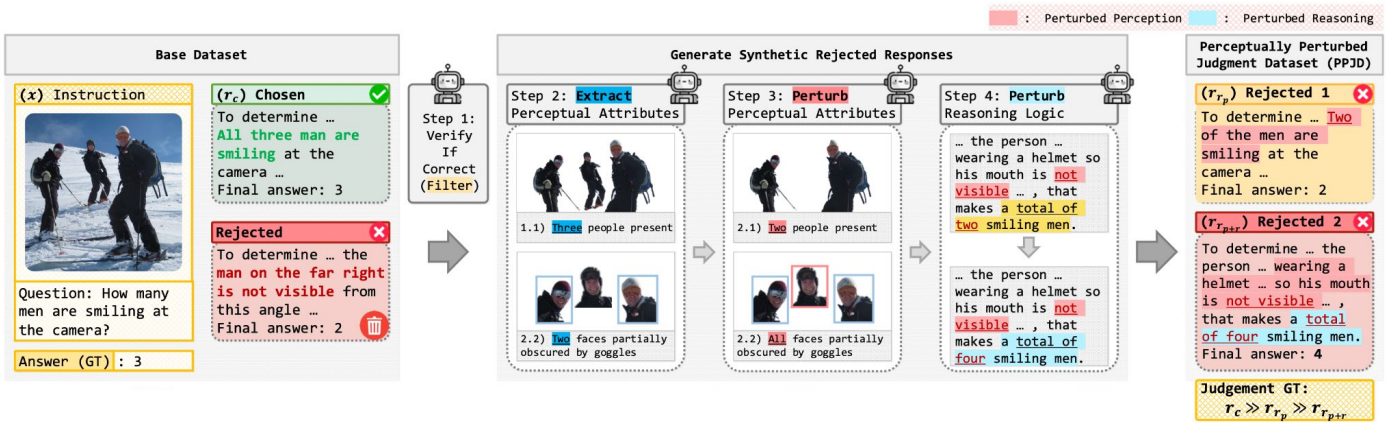
r_c vs. r_p — Contains visual error only. Reasoning remains fluent and plausible.



Conclusion: Existing judges reject bad reasoning, but miss fluent visual errors.

Dataset Generation -

PPJD: Perceptually Perturbed Judgment Dataset



For each input:

- r_c : perceptually & logically correct response
- r_{r_p} : perceptually incorrect but reasoning-preserved
- $r_{r_{p+r}}$: both perceptually and logically incorrect

Supervision

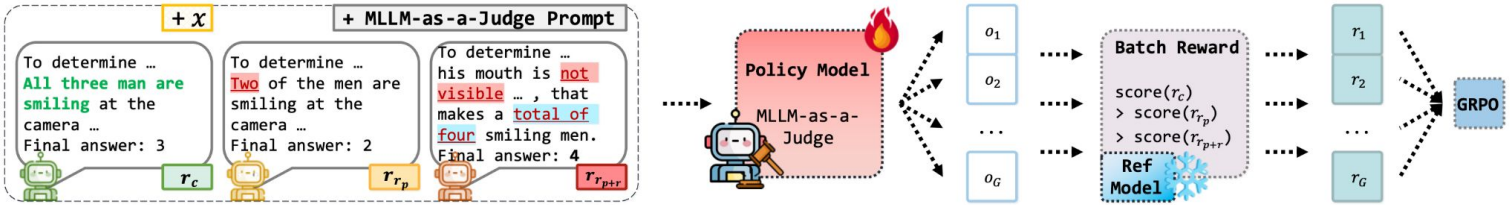
- Explicit ordering: $r_c \succ r_{r_p} \succ r_{r_{p+r}}$

PPJD Statistics

Table S1. Category-level statistics of the Perceptually Perturbed Judgment Dataset (PPJD).

Category	Included Datasets	Size	(%)
General VQA	VSR, TallyQA, OKVQA, GQA	1.7 K	45 %
Science	M^3CoT	0.6 K	15 %
Mathematics	MAVIS, GeomVerse	0.5 K	14 %
OCR	SROIE	0.5 K	14 %
Chart	MapQA	0.3 K	7 %
Image Quality	KonIQ-10k	0.2 K	5 %

Training Objective -
Batch-Ranking Reward with GRPO



Goal & Optimization

Train the judge to output the global target order. We use GRPO to optimize the judge with a verifiable reward. ($r_c \succ r_{r_p} \succ r_{r_{p+r}}$)

Reward Design

Format reward: Validates the judgment output structure.

Batch-ranking reward: Measures closeness to the target order (using Levenshtein distance).

Core Contribution: A perception-aware judge trained not with human pair labels, but with **verifiable structured perturbations.**

Result in MLLM-as-a-Judge Benchmark

Model	Single Score ↑	Pairwise (w/ Tie) ↑	Pairwise (w/o Tie) ↑	Batch Ranking ↓
Flex-Judge-VL-7B (Baseline)	0.404	0.514	0.623	0.517
Perception-Judge-Flex (Ours)	**0.466**	**0.520**	**0.645**	**0.505**
Qwen3-VL-4B-Thinking (Baseline)	0.419	0.543	0.663	0.498
Perception-Judge-Flex (Ours)	**0.457**	**0.554**	**0.691**	**0.444**

> Our model achieves state-of-the-art evaluation alignment, matching proprietary models like GPT-4o while remaining highly data-efficient (trained on just 3k samples).

MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark (ICML 2024 Oral)

ICML
International Conference
On Machine Learning

KAIST CVML

Thank you!

Paper

Project Page

Code