When Self-Preference Is a Verification Failure: Causal Evidence from Vision-Language Judges

Anonymous authors
Paper under double-blind review

Abstract

Vision-language models used as judges tend to prefer their own answers, and the usual explanation is narcissism: the judge recognizes its own output and rewards it. We argue instead that self-preference is mostly a verification failure: a judge checks each answer against its own interpretation of the image, so when it misinterprets the image, it confirms the wrong answer produced from that misunderstanding. We test this causally in 19 open vision-language judges from 8 model families, across five multimodal benchmarks and a text control, by changing the judge's image while keeping every candidate answer byte-identical. An outcome-matched control shows that most raw self-preference is evaluator noise, consistent with recent findings in text-only settings. The remaining bias follows the judge's interpretation of the image rather than authorship: answers are favored for the scene they were written from rather than because the judge wrote them. Degrading the judge's view steadily reduces this bias, and showing the judge the image used to generate its other answer reverses this preference. A judge's accuracy at verifying other models' answers on items it answers incorrectly predicts its favoritism, including for judges and families excluded from fitting; the signal comes from answers that repeat the judge's own mistake. Tests on nine larger judges, four commercial, show that this relationship predicts their much higher favoritism, but an authorship effect can no longer be ruled out: individual larger judges show substantial direct authorship preference, the strongest almost matching that judge's evidence effect. This mechanism has a practical cost and a diagnostic use: a model choosing among its own samples selects a correct answer less often than other judges do on average, and removing or replacing the image during evaluation changes which judges are identified as biased.

1 Introduction

Evaluation has quietly changed hands. Benchmarks, leaderboards and training pipelines now use one model to score another (Zheng et al., 2023), and vision-language models (VLMs) have joined the bench, grading answers about images at a scale no human panel could match (Chen et al., 2024a; Lee et al., 2024). Increasingly the loop closes on itself: a model ranks its own samples (Kang et al., 2025) and scores its own training data (Yuan et al., 2024). Each pipeline puts the judge's own answer on the ballot, so whether it can be trusted there decides whether a widely deployed practice is sound.

Research has gradually narrowed the evidence for self-preference driven by self-recognition. Panickssery et al. (2024) gave the worry its modern form: judges favor their own generations, the effect tracks their ability to recognize their own text, and it earned the name model narcissism. Wataoka et al. (2024) offered a quieter suspect, familiarity for text that is likely under the judge's own distribution. Later work isolated the harmful case, a judge preferring its own *inferior* answer (Chen et al., 2025). Then came the sanity check: Roytburg et al. (2026) showed with an outcome-matched control that about ninety percent of measured self-preference in text judges is evaluator noise. The indictment collapsed, but not to zero, leaving a residual with no explanation and a question: if it is not recognition, and self-recognition is itself weak (Davidson et al., 2024; Laine et al., 2024), what is left?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

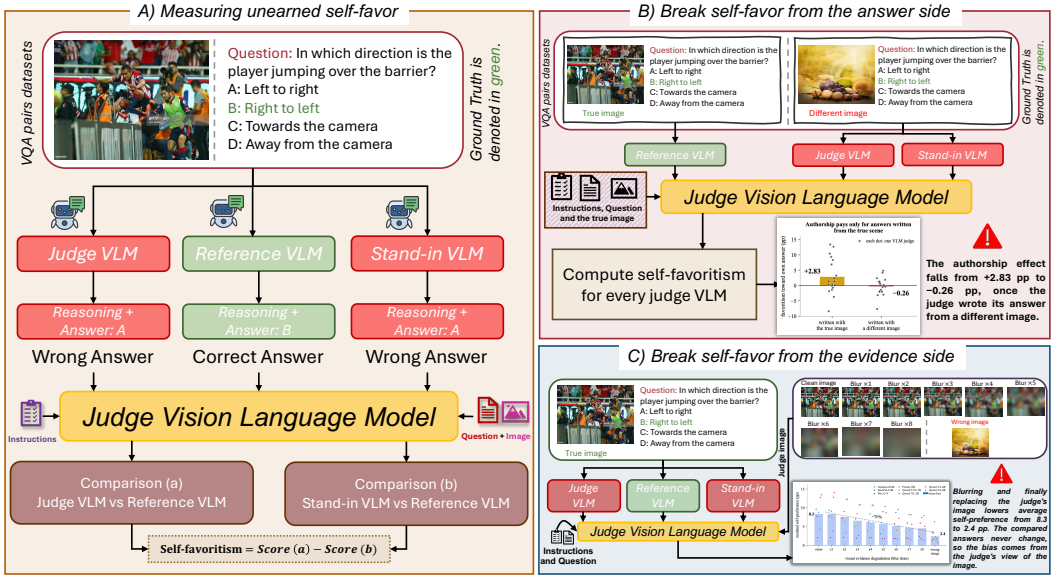


Figure 1: Overview on a real benchmark item. *(a)* The measurement: the judge compares its own wrong answer against a reference model's correct answer, with verdicts read from token probabilities. *(b)* Breaking the answer side: the judge's own answer earns $+2.83$ pp written with the true image in view and -0.26 pp written from a different scene. *(c)* Breaking the evidence side, by degrading the image at judging time, reduces what is left, partly through a generic loss of decisiveness, and every self-preferring judge declines.

While that debate ran in text, a second literature grew around multimodal judges and kept finding one observation: the judge does not really use the image. VLM judges prefer answers that merely sound informative when the image contradicts them (Zou et al., 2026), fail to integrate visual and textual evidence (Lee et al., 2026), are steerable by adversarial image edits (Hwang et al., 2025), and trail their own generation ability as reward models (Li et al., 2025b). The first large multimodal measurement found self-preference in captioning judges but offered no mechanism (Koyama et al., 2026). This paper shows that the two threads describe one phenomenon.

Our claim is that residual self-preference in VLM judges is grounded self-confirmation. When judging, the model does not so much recognize its answer as re-derive it. The judge reads the image, forms its own opinion of the answer, and scores each candidate by agreement with it. When the opinion is wrong, its own wrong answer agrees perfectly, so the judge confirms its own mistake while sincerely checking facts. Vision is not merely where this matters; it makes the mechanism testable: the image is a causal handle on the judge's evidence that leaves every response byte-identical, which no text-only manipulation offers.

Our main judges are open-weight and at most 12.4B parameters by design: the protocol reads graded verdicts from full token probabilities, which only open weights allow, and models of this size are widely deployed as local judges and reward models (Appendix A).

Across 19 open VLM judges from 8 model families, five multimodal benchmarks and a text control, plus a separate probe of nine larger judges, this paper contributes:

- **A causal explanation and its limits.** An outcome-matched control removes most of the raw effect as evaluator noise, and causal tests show that authorship does not explain the remaining bias. A cross-generation design separates who wrote an answer from the image used to generate it, showing that preference follows the image: the judge gives its own answer no advantage on average when that answer was generated from a different image. However, some larger judges directly prefer their own answers, and we cannot predict in advance which judges show this preference.

- **A mechanism that predicts.** Degrading only the image shown to the judge reduces the remaining bias for every cohort judge that shows it. A judge's accuracy in checking other models' answers to questions it gets wrong then predicts its favoritism, including for held-out judges, model families and larger judges; the signal comes from answers that repeat the judge's own mistake, so it tracks the same confirmation.
- **A cost and a diagnostic.** A model choosing among its own samples selects a correct answer less often than a panel of other judges would, sometimes less often than a random choice. Re-evaluating without the image or with another image removes much of the measured self-preference and changes which judges are flagged as biased.

2 Related Work

Model judges and their biases. Using one model to grade another's outputs is standard practice (Zheng et al., 2023) and now includes multimodal evaluation (Chen et al., 2024a; Lee et al., 2024), with documented biases involving answer position (Wang et al., 2024a) and cognitive biases (Koo et al., 2024; Chen et al., 2024b; Ye et al., 2025), which prompt changes only partially reduce (Soumik, 2026). These multimodal failures share a pattern: the judge does not really use the image. Evaluator VLMs fail to combine visual and textual evidence (Lee et al., 2026), perform worse as reward models than as generators (Li et al., 2025b), can be fooled by image manipulations (Hwang et al., 2025), and favor answers that sound informative over those supported by the image (Zou et al., 2026). This problem predates modern systems, appearing in language priors in VQA (Goyal et al., 2017) and continuing in today's models (Tong et al., 2024), motivating training with perceptual perturbations (Park et al., 2026). We show that self-preference is a visible sign of judges checking answers against their own, possibly mistaken, reading of the image.

Self-preference and self-recognition. Panickssery et al. (2024) linked self-preference to self-recognition; Wataoka et al. (2024) proposed familiarity instead. Related work showed that models strengthen their own biases when refining outputs (Xu et al., 2024), judges favor models trained on related data (Li et al., 2026), harmful self-preference can be identified using gold judgments (Chen et al., 2025), and self-preference occurs in multimodal captioning judges (Koyama et al., 2026). The recognition explanation is weak: models struggle to identify their own text (Davidson et al., 2024) and know little about themselves (Laine et al., 2024). Closest to our work, Roytburg et al. (2026) showed that an outcome-matched control removes about ninety percent of measured self-preference in text judges, leaving the residual unexplained. These studies are text-only or correlational. We extend the protocol to vision, causally establish that, on average, authorship does not explain the residual self-preference, and explain this remaining bias by changing the image.

The generation-verification gap and its stakes. Verification is increasingly viewed as a distinct capability (Lightman et al., 2024; Zhang et al., 2025), yet models struggle to check their own work: they cannot reliably self-correct (Huang et al., 2024), self-verification performs worse than external verification (Stechly et al., 2025), and the generation-verification gap remains a persistent model property (Song et al., 2025; Davidson et al., 2026). Our predictive relationship links verification to judge bias well enough to rank unseen judges and families. This matters because models increasingly filter outputs using their own judgments (Kang et al., 2025; Yuan et al., 2024): optimizing against a proxy evaluator risks overoptimization (Gao et al., 2023), and reinforcement learning from human feedback (RLHF) can teach models to be convincing rather than correct (Wen et al., 2025). We measure the cost of models selecting their own answers, explain why hiding authorship helps so little (Mahbub & Feng, 2026), and provide an ablation that measures the bias better.

3 Setup

Protocol and estimand. We measure only unearned self-preference. Each model M answers every item i once, producing response $o_M^{(i)}$ (a brief rationale plus a final answer line). An exact-match oracle assigns correctness $c\big(o_M^{(i)}\big) \in \{0, 1\}$, with 1 correct and 0 incorrect. Following the illegitimate self-preference (ILSP) protocol of Roytburg et al. (2026), judge J and reference R are

evaluated on

$$\mathcal{I}_{J,R} \;=\; \Big\{\, i \;:\; c\big(o_J^{(i)}\big) = 0,\;\; c\big(o_R^{(i)}\big) = 1,\;\; \exists\, K \notin \{J, R\} \text{ with } c\big(o_K^{(i)}\big) = 0 \,\Big\}, \tag{1}$$

items where the judge is wrong, the reference is correct, and another model K is also wrong. For each item, the judge sees the true image and compares its own wrong answer against the reference's correct answer, and each stand-in's wrong answer against that same reference answer. For responses a and b, let $s(a \succ b) \in [0, 1]$ denote the judge's preference score for a over b. Self-preference is

$$\mathrm{SF}_J \;=\; \mathbb{E}_{i \in \mathcal{I}_{J,R}}\Big[\, s\big(o_J^{(i)} \succ o_R^{(i)}\big) \;-\; \tfrac{1}{|\mathcal{K}_i|} \textstyle\sum_{K \in \mathcal{K}_i} s\big(o_K^{(i)} \succ o_R^{(i)}\big) \Big], \tag{2}$$

where $\mathcal{K}_i$ contains stand-ins wrong on item i, excluding J and R, and $|\mathcal{K}_i|$ is its size. The expectation averages over items within a (judge, reference, dataset) cell; cell scores then average into one SF_J per judge. Judges are the unit of inference for every headline test except the restyling control (Appendix G). The judge's and stand-ins' answers in Eq. 2 are wrong, so a fair judge treats them alike, giving $\mathrm{SF}_J = 0$: subtracting the stand-in average removes evaluator noise, leaving only unearned preference for the judge's own mistake.

Verdict readout. The judge sees the image, the question, and two responses labeled only as Assistant A and Assistant B; authorship is never revealed. We do not sample the verdict. At the decision position the next-token distribution assigns $P(\mathrm{A})$ and $P(\mathrm{B})$ to the verdict tokens, and we read the graded preference from them, presenting both orders and averaging:

$$s(a \succ b) \;=\; \tfrac{1}{2}\Big[\, p(a \,|\, a \text{ first}) \;+\; p(a \,|\, a \text{ second}) \,\Big], \qquad p(a \,|\, \cdot) \;=\; \frac{P(\ell_a)}{P(\mathrm{A}) + P(\mathrm{B})}, \tag{3}$$

where $\ell_a \in \{\mathrm{A}, \mathrm{B}\}$ is the letter labeling a. This removes position bias by construction (Wang et al., 2024a) and keeps the difference between a 55/45 lean and a 99/1 conviction that a sampled letter discards. A tie token is offered as well, and the mass on the verdict tokens is checked per judge (Appendix A).

Judges, benchmarks, scale. The cohort is 19 open-weight VLMs from 8 families spanning 0.9B to 12.4B parameters (Qwen, LLaVA, InternVL3, Phi, Gemma-3, Gemma-4, Pixtral, SmolVLM), four of which also serve as the fixed references; five size ladders inside three families support the scale analysis of Section 6. Checkpoints and the health gate are in Appendix A; MMMU and POPE enter the causal arms only for a pilot group of judges (Appendix D). We use five vision benchmarks and one text control, 500 items each, all with exact-match oracles: MMStar, MMMU, ScienceQA, MathVista, POPE and GSM8K, which has no images. Judges always see full free-text responses, and the final answer only labels correctness (free-form and open-ended checks: Appendix J). The study comprises about 1.5 million scored comparisons.

Interventions. Two manipulations carry the causal claims. *Cross-generation* breaks the answer side: the judge and the stand-ins re-answer item i while shown a different item's image (the donor image), and we keep only responses still wrong for item i, giving answers that are genuinely self-written yet not grounded in the true scene (Section 4). *Evidence degradation* breaks the image side: at judging time the image is blurred at up to eight log-spaced severities and finally replaced by another item's image at identical dimensions, so the visual token count never changes while the candidate responses stay byte-identical (Section 5). A no-image condition, two non-blur degradations, three alternative judge templates and a pointwise grading protocol serve as controls (Appendices F, B and C).

Statistical analysis. All primary findings survive Benjamini–Hochberg correction at a false discovery rate of 0.05. Uncertainty estimates are clustered by judge, and equivalence claims use two one-sided tests (TOST) with sensitivity to the margin (Appendix H).

4 Self-Preference Is Not Authorship

The prior result transfers to vision. Before going further than Roytburg et al. (2026), we replicate their result. Their central finding in text was that an outcome-matched control removes about

ninety percent of measured self-preference. The same holds for VLM judges: over our six benchmarks, the stand-in term of Eq. 2 removes 94.3% (89.6% in their text setting), from a raw preference of 0.537 for the judge's own wrong answer to a residual of 0.031, on exact-match oracles rather than LLM-derived gold labels. As in their analysis, the share is measured from zero, an accurate judge's score here: the control removes the 0.506 any wrong answer receives, which is evaluator noise, not favoritism. From coin-flip 0.5 instead, the raw excess is 3.7 pp, of which the residual is 3.1 pp. What remains is a small but real residual, and the rest of this paper is about it.

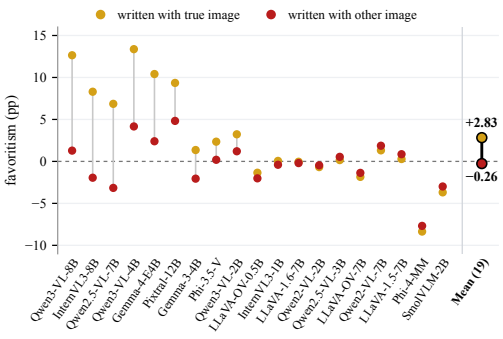
Design: separating authorship from image content. Narcissism and familiarity predict preference based on authorship; our mechanism predicts preference based on the image used to generate answers. The cross-generation 2×2 varies two factors: author (judge or stand-in) and generation image (true or different). For each illegitimate self-preference (ILSP) item, where the judge and at least one stand-in are wrong, the judge and stand-ins answer the same question with a different item's image, retaining answers $\tilde{o}$ only if still wrong for the original item. All candidates are wrong, the reference is correct, and judging always uses the true image. Authorship is never disclosed, so we test implicit preference rather than reactions to stated authorship. Three effects follow: A_{true} measures preference for the judge's own answers over stand-ins generated with the true image, equal to SF in Eq. 2; A_{alt} measures this preference for different-image answers $\tilde{o}$ instead of answers o; and $E_{\text{self}} = s(o_J \succ o_R) - s(\tilde{o}_J \succ o_R)$ compares the judge's own answers across images against the same reference. Recognition predicts $A_{\text{alt}} \approx A_{\text{true}} > 0$ regardless of image; grounded self-confirmation predicts $A_{\text{true}} > 0$ for true-image answers, but $A_{\text{alt}} \approx 0$.

Result: favoritism follows the image, not the author. Table 12 (Appendix H) and Figure 2a show the results. The interaction distinguishes the two proposed explanations: it is $+3.09$ pp ($p = 0.007$), positive in 12 of 19 judges, and unchanged when tested using a cluster bootstrap over judges or a mixed-effects model with random slopes (Table 13). The image used to generate an answer has a large effect: the judge scores its own true-image answer $+6.80$ pp higher than its different-image answer (both against the same reference answer) ($p = 0.003$). The authorship advantage is smaller: $+2.83$ pp for answers generated with the true image ($p = 0.048$; Wilcoxon $p = 0.066$). For answers generated with a different image, this advantage disappears, at -0.26 pp, with TOST placing the effect within ±1.5 pp ($p = 0.037$); its 90% interval, $[-1.39, +0.87]$ pp, excludes an advantage of 1 pp but not a penalty of that size. Thus, the image used to generate an answer matters much more than who wrote it. This small average authorship effect applies only to items the judge answered incorrectly, where self-preference is harmful, and describes the 19 judges collectively rather than every judge individually: four judges show small positive authorship effects on different-image answers (Appendix H), and on free-form answers a small effect cannot be ruled out (Appendix J). Section 6 examines larger judges.

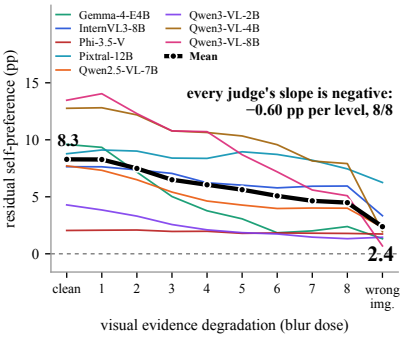
Style is not the channel. Because changing the image used to write an answer also changes its wording, we test style directly, the channel that authorship-obfuscation defenses remove through paraphrasing (Mahbub & Feng, 2026). Rewriting six self-preferring judges' own answers with a different model while preserving their content and correctness removes only about 10% of the effect: $+9.55$ pp remains. Replacing the image while keeping the rewritten text unchanged reduces the same effect to 0.84 pp in these evaluations. Authorship is still recoverable: a lexical classifier identifies which of the 19 models wrote a held-out response far above chance, but a model's recognizability does not predict its authorship effect on different-image answers ($r = +0.17$). Controls for likelihood and length also leave the overall effect intact (Appendix G).

5 Grounded Self-Confirmation

The mechanism. We explain the residual through grounded self-confirmation: the judge re-derives its own opinion of the correct answer from the image and scores each candidate by agreement with it. On ILSP items this opinion is wrong and the judge's own answer agrees with it perfectly, so the judge confirms its own mistake while sincerely checking the facts. A direct probe of this re-derivation supports it only in part (Appendix O). This predicts what the earlier accounts do not: the bias should fall as the judge's visual evidence weakens.



(a) Breaking the *answer* side: per-judge favoritism for own answers written with the true image (gold) or a different image (red), judged with the true image.



(b) Breaking the *evidence* side: residual self-preference as the judging image is blurred, then replaced. Bold: mean of the 8 judges with measurable bias (all 19: Table 4).

Figure 2: The two causal manipulations: (a) changing the scene an answer was written from, and (b) degrading the judge's view at judging time.

Dose-response. We re-evaluate the same comparisons while changing only the image shown during judging. We use eight log-spaced blur levels between the clean image and a mismatched image, while every candidate response remains byte-identical across all ten conditions. For the eight judges with measurable baseline bias (Figure 2b), the mean residual drops from 8.3 pp to 2.4 pp across the ten levels. This decline is consistent across judges, not only in the average: every judge has a negative slope, averaging -0.60 pp per level (95% CI $[-0.97, -0.22]$, $p = 0.007$, 8 of 8; Appendix H). The result also holds across all 19 judges: the clean-to-mismatch drop is $+2.57$ pp ($p = 0.012$), positive for 15 of 19.

The image degradation might simply make all verdicts less decisive, moving them toward indifference. We therefore estimate this ordinary shrinkage from stand-in comparisons; after removing it, the judge's own scores still decrease by an additional -0.9 to -1.3 pp under both calibrations, showing that a self-specific effect remains (Appendix G).

The preference reverses, which shrinkage cannot do. The judge writes two answers: one with the true image and one with a donor image. Re-evaluating them with the true image, it prefers the true-image answer by $+6.80$ pp. Showing the donor image favors the donor-image answer, changing the effect to -5.10 pp; the shift is negative for 13 of 19 judges ($p = 0.007$). The authorship contrasts also switch. This shows that the judge favors the answer generated from the image it is currently viewing, rather than simply favoring the answer it wrote. The judge wrote both answers, so familiarity cannot explain it; shrinkage cannot either: it can only move scores toward indifference, not change their sign (Appendix E).

It is not blur, not our prompt, and not the pairwise format. Pixelation and additive noise reduce the bias in the same direction as blur (Appendix F), three judge templates reproduce the decline (Appendix B), and grading one answer at a time makes self-favoritism larger, not smaller ($+7.49$ versus $+3.18$ pp), and it still collapses with a swapped image (Appendix C).

6 A Predictive Law

From mechanism to forecast. Grounded self-confirmation predicts that judges favor their own mistakes more when they are less able to detect them. We therefore measure verification accuracy on ILSP items, where the judge's original answer is wrong, by asking it to compare other models' wrong answers with the reference's correct answer:

$$\mathrm{VA}_J \;=\; \mathbb{E}_{i,\,K\in\mathcal{K}_i}\Big[\,1 - s\big(o_K^{(i)} \succ o_R^{(i)}\big)\Big], \tag{4}$$

where the expectation averages over eligible items i and their incorrect stand-ins $K \in \mathcal{K}_i$. Since s here is preference for the wrong answer, VA_J is a graded preference for the correct one (higher is better verification), not a fraction of sampled verdicts, and it uses only stand-in comparisons, never the judge's own answer. We use VA_J to predict self-preference SF_J. To prevent shared measurement noise from creating a relationship between them, we estimate VA_J on one half of the items and SF_J on the other, then swap the halves and average the results. This procedure is called cross-fitting (Appendix K).

One variable explains differences between judges. We predict self-preference from verification accuracy using a linear model across judges,

$$\mathrm{SF}_J \;=\; \alpha + \beta\,\mathrm{VA}_J + \varepsilon_J, \qquad \hat{\beta} = -1.86 \;\; (p < 10^{-4}), \tag{5}$$

where ε_J is judge J's error; the negative slope means that judges who verify better show less self-preference. Across all 19 judges (Figure 3a), the correlation is $r = -0.850$ ($p = 4 \times 10^{-6}$), and verification accuracy alone explains 72.3% of the variation in self-preference between judges. Adding generation accuracy increases the explained variation by only three percentage points, without a significant improvement ($p = 0.15$). The plot also shows that 16 of the 19 judges fall below chance-level verification accuracy (0.5), and only Phi-4-MM clearly exceeds it, with greater self-preference the farther a judge falls below chance.

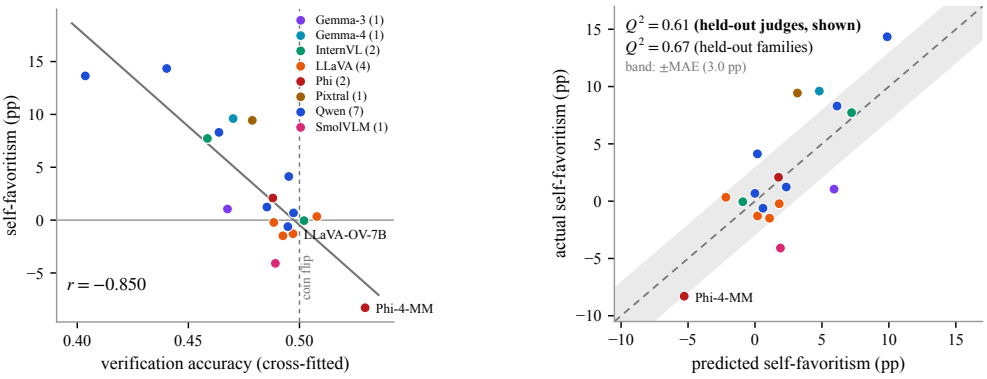
We also test the model on judges excluded from fitting, scoring it by Q^2 (Eq. 6 in Appendix K; zero means no better than predicting the mean). Leaving out each judge in turn gives $Q^2 = 0.61$, with a mean error of 3.0 pp compared with 5.2 pp for predicting the cohort mean. Leaving out each entire family, so no judge in the fit shares the held-out family's training recipe, gives $Q^2 = 0.67$ and a smaller error of 2.7 pp (Figure 3b). Generation accuracy alone predicts unseen families worse than the cohort mean (Appendix K). Finally, shared scores do not produce it: computed from cells with different reference models, which share no answers, it is no weaker than within one reference (Appendix K).

The predictor is the judge confirming its own mistake. It is not general skill: on a common set of 4,128 comparisons graded by all 19 judges, verification accuracy correlates positively with favoritism ($r = +0.62$), and it opposes the error-item measure ($r = -0.58$; Appendix L). On the judge's error items, about half of the stand-ins give the judge's exact wrong answer, and accuracy on those alone carries the whole relationship ($r = -0.89$). On stand-ins that are wrong differently, every judge sits near chance (0.48 to 0.51), judges do not reliably differ, and accuracy does not track favoritism ($r = -0.06$). The predictor measures how readily a judge endorses its own mistake in another model's words: the confirmation behind its favoritism, observed without its own answers. It is a consistency relationship and a diagnostic, not evidence that a checking skill causes the bias (Appendix K).

The exceptions are explained. Phi-4-MM, the cohort's one strongly negative judge (-8.3 pp), has the highest verification accuracy of all 19 (0.530): superior self-criticism, not inverted narcissism, and the law predicts its sign and most of its size. The LLaVA family's flat zero has an independent explanation. Judging-side grounding, the average shift in a judge's verdicts when its image is swapped, is 0.025 for LLaVA against 0.085 for the other 15 ($p = 0.001$): they barely utilize the image while judging, so there is no reading of the scene to confirm (Appendix K).

Scale widens the gap. In all five size ladders of our cohort, larger models answer more accurately without a matching gain in verification. The gap between generation and verification accuracy therefore widens in all five series, while self-preference increases in four. Parameter count alone does not predict self-preference ($\rho = 0.35$, $p = 0.14$), but the gap does: judges whose generation outpaces their verification show more bias.

Where the explanation is incomplete. All judges used to fit the model have at most 12.4B parameters, so we tested nine larger ones, four commercial (Appendix M). Applied unchanged, the model predicts their self-favoritism with $Q^2 = 0.87$ against the cohort mean, capturing their level but not their order. However, authorship sometimes matters. As a group, the nine resemble the original cohort: their mean A_{alt} is $+4.10$ pp, which this sample size cannot distinguish from zero



(a) Verification accuracy versus self-favoritism, one point per judge; the dashed line marks coin-flip verification.

(b) Each judge's favoritism predicted with that judge held out, against its actual value. Band: ± mean absolute error.

Figure 3: The predictive law. Verification accuracy on the judge's error items, measured on others' answers only, explains (a) and forecasts (b) self-favoritism.

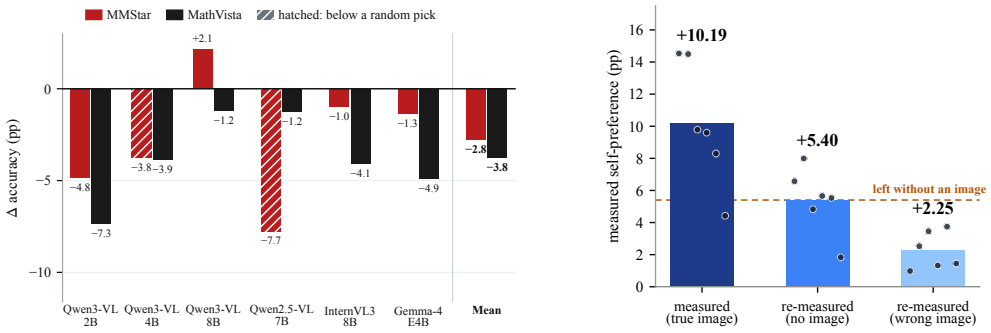
($p = 0.18$). Individual judges differ: GPT-4.1-mini favors its own answer almost equally when the image contradicts it and when it supports it (+21.79 versus an A_{true} of +22.34), which verification failure alone cannot explain. Some judges above this scale therefore directly favor their own answers, but we cannot predict which ones. Sensitivity to visual evidence separates them within the nine, but not across all 28 judges, and neither do the other measures we tested (Appendix M).

What carries to text. The same protocol on GSM8K identifies what depends on modality. Self-preference exists in text (+2.67 pp), and more self-favoring multimodal judges also tend to favor themselves in text ($r = +0.71$). However, verification accuracy predicts favoritism negatively only in multimodal tasks: in text, better generators are also better verifiers. The generation-minus-verification gap remains predictive in both ($r = +0.50$; Appendix I).

7 Consequences and an Audit Correction

Self-judging reduces best-of-N accuracy. The bias matters when models select among their own outputs. For six generators from three families, we sample seven answers per item and keep items where the answers disagree. With the true image shown, a judge compares every pair to select a winner. All six models judge every answer set, allowing comparison with a panel of other judges. Self-judging lowers accuracy by 3.40 points per generator relative to the panel mean ($p = 0.012$), performs worse in 11 of 12 generator-dataset cells (Figure 4a), and the disadvantage remains after controlling for generator, judge and dataset effects. It performs below random selection twice and below the generator's greedy answer in 5 of 12 cells; the panel average never falls below random selection. The panel judges are not better overall (Appendix N); their errors plausibly do not follow the generator's image misinterpretations, as our mechanism predicts but a narcissism explanation does not.

Re-measuring with the image removed or replaced. The mechanism also suggests a partial correction for measuring bias: we repeat the same comparisons with the image removed or replaced by another item's image, with no new labels, retraining or prompt changes, and what remains is the bias that survives without the true scene. Labels are still needed to select eligible items, so only this control, not the whole audit, is label-free. Figure 4b shows results for the six most self-preferring judges with complete coverage: the original +10.19 pp falls to +5.40 without the image and +2.25 with a wrong image. Withholding the image removes 47%: with no image, as under the heaviest blur, the judge can still read the question, and its own answer tends to agree with that reading. A wrong image removes 78%, because the judge's reading then follows the scene it is shown. The two differ significantly ($p = 0.019$), so we report both (Appendix F). Neither method creates bias where



(a) Best-of-7 accuracy lost to self-judging: the generator's own selection minus the mean of the five other judges on the same candidates. Hatched bars fall below random selection.

(b) Self-preference measured with the true image, no image and a wrong image. Dots: the six most self-preferring judges; bars: their mean.

Figure 4: The cost of self-judging (a) and the image-ablated re-measurement (b).

none exists: across all 18 judges with no-image coverage, the measured $+2.78$ pp falls to $+0.55$ without the image and $+0.02$ with a wrong image, neither distinguishable from zero.

Raw scores identify different models as most biased. Raw and image-adjusted self-preference give similar overall rankings (Spearman $\rho = 0.879$), making the raw measure appear reliable. However, the highest-ranked judges, the ones an audit would flag, change: Qwen3-VL-8B ranks first raw ($+15.1$ pp) but sixth when re-measured with a wrong image ($+1.5$ pp), while Pixtral-12B moves to first (Appendix N).

8 Conclusion

Across 19 open VLM judges, 8 families and about 1.5 million judgments, self-preference in judges that read the scene is a verification failure rather than vanity. Outcome matching removes most of the raw effect as evaluator noise, and what remains is not authorship: on average, an answer earns nothing if it was written from a different scene. The residual depends on the judge's own reading of the scene: degrading the image drives it down, and swapping the image reverses which of its own answers the judge prefers. How well a judge verifies others' answers where it makes mistakes then predicts which judges are biased, on judges and families the fit never saw; the signal comes from answers that repeat its own mistake. Some larger judges also show a direct authorship channel, which this account does not cover and we cannot yet predict. In our size ladders, generation grows faster than verification, so the gap may keep widening. Until training closes it, evaluation pipelines should assume that models grade the world the way they misread it.

AI use statement

In this work, we used a generative AI coding assistant (Claude, Anthropic) for experiment orchestration, figure generation, and editing the paper text under the direction of the authors, who specified the hypotheses, the experimental designs, and the claim structure. All AI-drafted and AI-edited text was reviewed and edited by the authors. All quantitative claims were checked against a frozen results file produced by the analysis pipeline. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

This work studies evaluation biases of publicly released vision-language models on public benchmarks and involves no human subjects or private data. Its findings caution against unaudited model-based evaluation, which we believe reduces rather than raises the risk of harmful deployment.

Reproducibility statement

All datasets are public. Every judge behind the main results is open-weight; the scale probe of Appendix M additionally reports four commercial judges reached through vendor APIs, recorded there by exact model string and access date. The appendix specifies the judging protocol, prompts, decoding settings, and perturbation parameters. The full pipeline code, the frozen results file from which every number in the paper is drawn, and per-judge result tables will be publicly available upon acceptance.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv preprint arXiv:2404.14219*, 2024.

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, et al. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. *arXiv preprint arXiv:2503.01743*, 2025.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, et al. Pixtral 12B. *arXiv preprint arXiv:2410.07073*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing Multimodal LLM-as-a-Judge with Vision-Language Benchmark. In *International Conference on Machine Learning (ICML)*, 2024a.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the Judge? A Study on Judgement Bias. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024b.

Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the Surface: Measuring Self-Preference in LLM Judgments. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Tim R. Davidson, Viacheslav Surkov, Veniamin Veselovsky, Giuseppe Russo, Robert West, and Caglar Gulcehre. Self-Recognition in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Tim R. Davidson, Anja Surina, and Caglar Gulcehre. The Future of Facts: Tracing the Factual Generation-Verification Gap. *arXiv preprint arXiv:2605.27564*, 2026.

Leo Gao, John Schulman, and Jacob Hilton. Scaling Laws for Reward Model Overoptimization. In *International Conference on Machine Learning (ICML)*, 2023.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, et al. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 model card, 2026. URL `https://huggingface.co/google/gemma-4-E4B-it`.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. In *International Conference on Learning Representations (ICLR)*, 2024.

Yerin Hwang, Dongryeol Lee, Kyungmin Min, Taegwan Kang, Yongil Kim, and Kyomin Jung. Fooling the LVLM Judges: Visual Biases in LVLM-Based Evaluation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2025.

Zhewei Kang, Xuandong Zhao, and Dawn Song. Scalable Best-of-N Selection for Large Language Models via Self-Certainty. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking Cognitive Biases in Large Language Models as Evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Shuitsu Koyama, Yuiga Wada, Daichi Yashima, and Komei Sugiura. MLLM-as-a-Judge Exhibits Model Preference Bias. *arXiv preprint arXiv:2604.11589*, 2026.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Jeremy Scheurer, Mikita Balesni, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. Prometheus-Vision: Vision-Language Model as a Judge for Fine-Grained Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.

Sua Lee, Sanghee Park, and Jinbae Im. MM-JudgeBias: A Benchmark for Evaluating Compositional Biases in MLLM-as-a-Judge. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research (TMLR)*, 2025a.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference Leakage: A Contamination Problem in LLM-as-a-judge. In *International Conference on Learning Representations (ICLR)*, 2026.

Lei Li, Yuancheng Wei, Zhihui Xie, Xuqing Yang, Yifan Song, Peiyi Wang, Chenxin An, Tianyu Liu, Sujian Li, Bill Yuchen Lin, Lingpeng Kong, and Qi Liu. VL-RewardBench: A Challenging Benchmark for Vision-Language Generative Reward Models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations (ICLR)*, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, 2024b.

Taslim Mahbub and Shi Feng. Mitigating Self-Preference by Authorship Obfuscation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 37701–37708, 2026.

Andrés Marafioti, Orr Zohar, Miquel Farré, Merve Noyan, Elie Bakouch, et al. SmolVLM: Redefining small and efficient multimodal models. *arXiv preprint arXiv:2504.05299*, 2025.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM Evaluators Recognize and Favor Their Own Generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Seojeong Park, Jiho Choi, Junyong Kang, Seonho Lee, Jaeyo Shin, and Hyunjung Shim. Mitigating Perceptual Judgment Bias in Multimodal LLM-as-a-Judge via Perceptual Perturbation and Reward Modeling. In *International Conference on Machine Learning (ICML)*, 2026.

Dani Roytburg, Matthew Bozoukov, Matthew Nguyen, Jou Barzdukas, Mackenzie Puig-Hall, and Narmeen Oozeer. Are LLM Evaluators Really Narcissists? Sanity Checking Self-Preference Evaluations. In *International Conference on Machine Learning (ICML)*, 2026.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham Kakade, Dean Foster, and Udaya Ghai. Mind the Gap: Examining the Self-Improvement Capabilities of Large Language Models. In *International Conference on Learning Representations (ICLR)*, 2025.

Sadman Kabir Soumik. Judging the Judges: A Systematic Evaluation of Bias Mitigation Strategies in LLM-as-a-Judge Pipelines. *Transactions on Machine Learning Research (TMLR)*, 2026.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. On the Self-Verification Limitations of Large Language Models on Reasoning and Planning Tasks. In *International Conference on Learning Representations (ICLR)*, 2025.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large Language Models are not Fair Evaluators. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, et al. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-Preference Bias in LLM-as-a-Judge. In *NeurIPS 2024 Workshop on Safe Generative AI*, 2024.

Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. Language Models Learn to Mislead Humans via RLHF. In *International Conference on Learning Representations (ICLR)*, 2025.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. In *International Conference on Learning Representations (ICLR)*, 2025.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-Rewarding Language Models. In *International Conference on Machine Learning (ICML)*, 2024.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. Generative Verifiers: Reward Modeling as Next-Token Prediction. In *International Conference on Learning Representations (ICLR)*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, et al. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025.

Xiaohan Zou, Roshan Sridhar, Mohammadtaher Safarzadeh, and Dan Roth. When Vision-Language Models Judge Without Seeing: Exposing Informativeness Bias. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026.