DO MLLM JUDGES JUDGE THE EDIT? AUDITING BIAS IN IMAGE EDITING EVALUATION WITH VERIFIED QUALITY PRESERVATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) are increasingly used as automated judges for instruction-based image editing and as reward signals for model training. However, systematically auditing whether these judges are influenced by cues irrelevant to editing quality is challenging because visual interventions may themselves alter the quality being evaluated. A judgment shift can therefore be attributed to bias only when the intervention is verified to preserve the underlying editing quality. To address this challenge, we introduce EditJudgeBias, a counterfactual benchmark with verified quality preservation, comprising 1,196 real editing samples and 13 cues injected across four evaluation sites. We verify quality preservation for the requested edit using calibrated multimodal validators, controls, and human inspection. We then audit five MLLM judges along three complementary dimensions: invariance to quality-preserving cues, agreement with human judgments, and stability of pairwise preferences. Importantly, observed shifts are evaluated against each judge's own zero-dose and re-query noise floors rather than against zero. Experiments show that quality-preserving cues move every judge beyond its own noise. Fabricated majority opinions increase ratings, irrelevant visual elements cause larger shifts than whole-image manipulations, and swapping candidate order reverses up to 60.9% of pairwise decisions. Edit-region cues also tend to reduce human agreement. The three measures characterize judges differently, showing that robustness cannot be captured by a single metric.

1 INTRODUCTION

Recent advances in unified multimodal models have substantially improved instruction-based image editing (Deng et al., 2025; Xie et al., 2025; Diao et al., 2026; Fu et al., 2026). Evaluating these models requires judging instruction adherence, preservation of unrelated content, and visual quality. Because human evaluation is costly, benchmarks increasingly use multimodal models as automated judges (Zhao et al., 2025; Wu et al., 2025a), while dedicated evaluators are trained on human preferences (Wu et al., 2026; Ji et al., 2026). These judges are also used for output selection and reinforcement learning (Luo et al., 2026), making their sensitivity to the evaluation procedure consequential. Do they measure the edit, or do quality-preserving cues move their decisions?

Agreement with human judgments provides only a partial view of judge reliability. A judge may align well with humans on the original benchmark yet change its decision when the same edit is presented differently. Attributing an output to a particular model or claiming that other evaluators prefer it, for example, provides no new evidence about whether the instruction was followed. Visual cues are harder to test because the intervention itself may affect the edit: a region annotation can obscure an object, and a brightness adjustment can change an attribute requested by the instruction. Figure 1 illustrates why preservation has to be checked as part of the audit.

We introduce **EditJudgeBias**, a counterfactual benchmark for testing these sensitivities under controlled interventions. It contains 1,196 real editing triplets drawn from five public benchmarks and three independent content pools, with thirteen cues inserted at four points in the evaluation pipeline: the display protocol, judge prompt, image pixels, and edited content. Each sample is judged both

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: **Bias-relevant cues occur naturally in image-editing benchmarks.** Four real examples illustrate framing, typographic, provenance, and colorfulness cues. Human ratings may not penalize these cues, yet MLLM judges explicitly attend to them; EditJudgeBias tests these effects under controlled interventions

with and without the cue. We collect three 1–10 rubric ratings in the scoring setting and preferences between two edits of the same request in the pairwise setting. Image-side cues are checked with calibrated validators and human annotations, and observed shifts are interpreted relative to each judge's SHAM and re-query variability.

Quality-preserving cues systematically affect all five judges beyond their own response noise. In scoring, a fabricated majority opinion raises all three ratings across all judges, while a small sticker causes larger shifts than any whole-image manipulation. These effects also extend to human agreement: when a cue significantly changes agreement, it almost always reduces it, particularly for cues involving the edit region. In pairwise evaluation, simply swapping the display order reverses up to 60.9% of decisions, far exceeding the 1.0 to 10.5% reversal rate under repeated queries. Moreover, the judges behave differently across invariance, agreement, and stability, suggesting that no single robustness measure captures all failure modes.

2 Related Work

2.1 Judges for Instruction-Based Image Editing

Instruction-based image editing (Brooks et al., 2023) is commonly evaluated on benchmarks that pair edited images with human labels (Zhang et al., 2023; Ku et al., 2024b; Jiang et al., 2024; Ma et al., 2024; Xu et al., 2025). VIEScore helped establish rubric-prompted MLLM scoring for this setting (Ku et al., 2024a), an approach now used in leaderboards (Liu et al., 2025a; Ye et al., 2025b), agentic evaluators (Wang et al., 2025a), and open judges (Lee et al., 2024). A parallel line of work trains editing-specific reward models against human preferences and uses them for online reinforcement learning (Luo et al., 2026; Zhao et al., 2026; Chen et al., 2026a; Xu et al., 2026; Wang et al., 2025b). Other studies compare judges with human labels through factor-level agreement (Liu et al., 2026), expert mean-opinion scores and intermediate-layer probing (Gao et al., 2026; Sun et al., 2026), or dedicated preference suites (Bai et al., 2026).

Sensitivity to quality-preserving changes in evaluation has received much less attention. EditReward reports a position effect for one judge in an appendix (Wu et al., 2026); GEditBench v2 randomizes presentation order without quantifying the resulting position bias (Jiang et al., 2026); and

EdiVal-Agent catalogs failure modes of zero-shot judges without controlled interventions (Chen et al., 2026c). They leave open whether presentation changes that preserve the underlying edit can move judge decisions beyond ordinary response variability. The issue also matters when judges are used as reward models, since optimization can amplify exploitable shortcuts (Gao et al., 2023).

2.2 Biases in MLLM-as-a-judge

Text-only judges are known to exhibit position, verbosity, and self-preference biases (Zheng et al., 2023; Wang et al., 2024; Saito et al., 2023; Panickssery et al., 2024). CALM systematically studies twelve such biases using robustness and consistency measures (Ye et al., 2025a). Recent work also shows that high consistency across repeated evaluations can coexist with substantial position bias (Norman et al., 2026), motivating construct validity as a framework for evaluating judges (Chen et al., 2026b).

Multimodal judges face additional biases arising from visual inputs. In text-to-image (T2I) evaluation, FRAME studies eight visual manipulations and shows that they can inflate judge scores (Hwang et al., 2025). Other studies investigate adversarial perturbations (Wang et al., 2026), preference bias, and prototypicality bias (Chen et al., 2025; Roy et al., 2026). In visual question answering and image captioning, MM-JudgeBias evaluates nine types of perturbations across 26 MLLMs (Lee et al., 2026), while other studies examine informativeness, perceptual, model-preference, and order biases (Zou et al., 2026; Park et al., 2026; Koyama et al., 2026; Ianaro et al., 2026). However, these audits focus on generated images, answers, or captions rather than instruction-based image editing. Evaluating an edit requires assessing whether the edited image faithfully follows the instruction given the source image, which introduces editing-specific factors such as the edit's location and extent. Cues tied to the edit region therefore remain largely unexplored in existing bias audits.

2.3 Quality verification of injected cues

Perturbation-based audits are interpretable only when the intervention leaves the evaluated construct unchanged; otherwise, a changed score may reflect a real degradation rather than evaluator bias. CheckList formalizes this as an invariance test (Ribeiro et al., 2020). For LLM judges, Chen et al. (2026b) report both invariance under construct-preserving edits and sensitivity to construct-changing ones, while Perturbation CheckLists deliberately degrade outputs to test the latter (Sai et al., 2021). Preservation is handled less uniformly elsewhere: it may be assumed (Vu et al., 2022), checked on a small manual sample (Wu et al., 2025b), left unverified for rewritten variants (Liu et al., 2025b), or avoided by collecting human ratings for both versions (Huang & Baldwin, 2023).

Multimodal audits make different choices as well. FRAME does not verify that its manipulations are harmless and lists semantic-preserving perturbations as future work (Hwang et al., 2025); MM-JudgeBias treats its composed perturbations as preservation-preserving (Lee et al., 2026); and Agarwal et al. (2026) verify 480 image–text items with three annotators each. Pixel or embedding similarity (Wang et al., 2004; Zhang et al., 2018; Hessel et al., 2021) measures how much an image changes, but not whether the edited output still satisfies its instruction. Even near-invisible overlaid text can alter model behavior (Cheng et al., 2024). We evaluate preservation relative to the source image and instruction, using calibrated validators and human checks for every image-side cue.

3 EditJudgeBias: Benchmark Design

Figure 2 summarizes the evaluation framework, benchmark composition, and human supervision used in EditJudgeBias.

3.1 Problem Formulation

Let $x = (I_{\text{src}}, ins, I_{\text{edit}})$ denote an editing sample comprising a source image, an editing instruction, and an edited image. Let $Q(x)$ denote its underlying quality across the three dimensions evaluated by the judges: instruction adherence, editing quality, and detail preservation. A judge J estimates Q under one of two protocols. In *scoring*, $J(x) \in [3, 30]$ is the sum of three 1–10 ratings, one per dimension. In *pairwise* evaluation, $J(x_1, x_2) \in \{\text{A}, \text{B}, \text{Tie}\}$ compares two edits of the same (I_{src}, ins), which is mainly dependent on I^a_{edit} and I^b_{edit}.

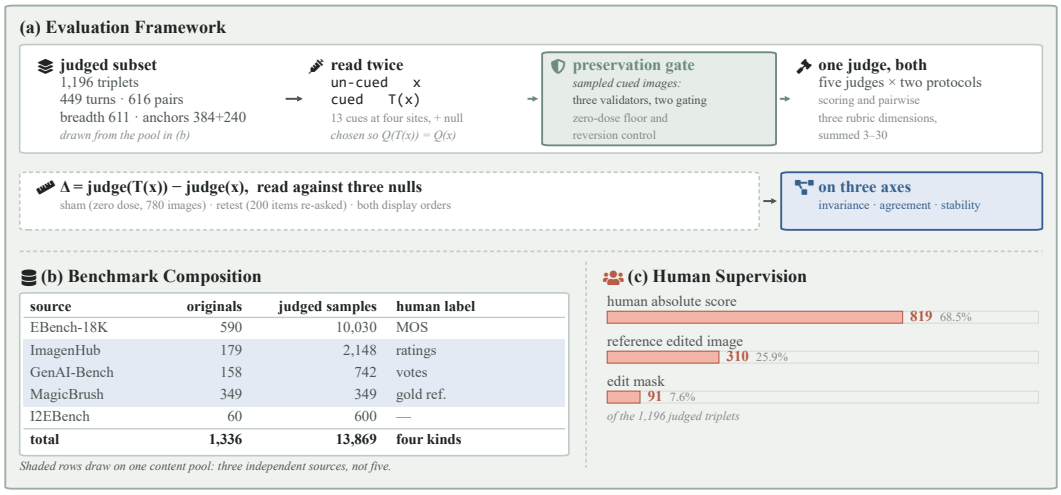

Figure 2: **Overview of EditJudgeBias.** (a) Evaluation framework for measuring judge invariance, agreement with human judgments, and stability under quality-preserving cues. (b) Composition of 1,196 editing triplets drawn from five public benchmarks spanning three independent content pools. (c) Human supervision used to assess agreement with human judgments and verify edit quality.

We define a *cue* as a transformation T of the edited image or the evaluation procedure that preserves the underlying quality, so that $Q(T(x)) = Q(x)$. A *bias* arises when the judge's decision depends on this cue despite the preservation of Q. We distinguish the magnitude of this dependence from its direction: cue-induced changes are evaluated against judge-specific response-noise baselines, while signed-shift tests separately assess whether paired changes have a consistent direction (Section 4).

For judge j, cue c, and rating dimension d, let $x_i^{(c)}$ and $x_i^{(0)}$ denote the cued and un-cued conditions, respectively. For the signed analysis, let $x_i^{(\mathrm{ref},c)}$ denote the cue-specific reference condition.

$$
\begin{aligned}
I_{j,c,d} &= \frac{10}{n^I_{j,c}} \sum_{i=1}^{n^I_{j,c}} \left| J_{j,d}(x_i^{(c)}) - J_{j,d}(x_i^{(0)}) \right|, \\
\Delta_{j,c,d} &= \frac{1}{n^\Delta_{j,c}} \sum_{i=1}^{n^\Delta_{j,c}} \left[J_{j,d}(x_i^{(c)}) - J_{j,d}(x_i^{(\mathrm{ref},c)}) \right].
\end{aligned}
\tag{1}
$$

A judge that faithfully measures Q should be invariant to T. In scoring, its ratings on all three dimensions should remain unchanged, implying $J(T(x)) = J(x)$. In pairwise evaluation, $J(T(x_1), x_2) = J(x_1, x_2)$; if T swaps the presentation order, $J(x_2, x_1)$ should return the corresponding mirrored preference. We assess *invariance* by the mean absolute per-item change in each rating dimension, *agreement* by the change in agreement with human judgments, and *stability* by preference preservation under display-order swaps.

3.2 Editing Samples

We build EditJudgeBias from five published image-editing benchmarks with human supervision: I2EBench (Ma et al., 2024), MagicBrush (Zhang et al., 2023), ImagenHub (Ku et al., 2024b), GenAI-Bench (Jiang et al., 2024), and EBench-18K (Xu et al., 2025). Although these are five benchmarks, they span only three independent content pools because the rated ImagenHub images are drawn from the MagicBrush development set and most GenAI-Bench editing prompts come from the same pool. Their annotations provide the supervision needed for different parts of the audit, including per-dimension mean-opinion scores, absolute ratings, pairwise preferences, and gold edits (Appendix B).

The selected samples are organized around two complementary uses. A *breadth* block of 611 samples, stratified by source and edit type, supports the invariance analysis across diverse edits. Two *anchor* blocks instead preserve complete editing turns, keeping multiple editors' outputs for the

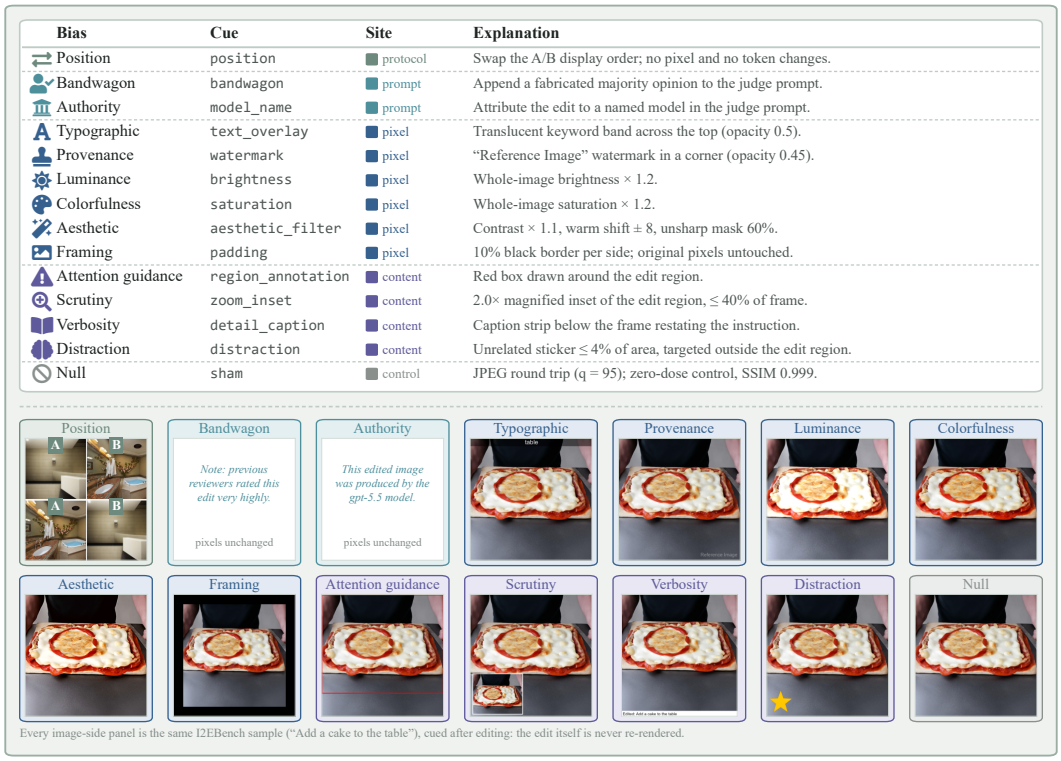

Figure 3: **Overview of the thirteen bias cues and the zero-dose control.** Each row identifies a cue, its injection site in the judging pipeline, and its intervention and dose. All image-side cues are illustrated using the same edited I2EBench example. POSITION is illustrated by the same judged pair under both display orders; it changes neither pixels nor prompt text.

same source image and instruction so that judge scores can be compared with human judgments. The anchor blocks contain 384 EBench-18K samples and 240 ImagenHub samples.

3.3 BIAS CUES

We introduce 13 controlled cues into the selected editing samples, grouped by where they enter the judging pipeline: protocol, prompt, pixel, or content as shown in Figure 3. This grouping identifies each intervention's entry point and helps localize where an observed effect may arise. Protocol and prompt cues introduce no visual changes and extend biases studied in text-only evaluation (Ye et al., 2025a; Panickssery et al., 2024). Pixel cues modify image pixels through photometric, framing, or localized overlay operations (Hendrycks & Dietterich, 2019; Cheng et al., 2024). Content cues are defined relative to the edit region and are specific to image-editing evaluation. Exact operators and doses are provided in Appendix B.

For content cues, we use the benchmark-provided source mask when available; otherwise, we estimate the edit region as the largest connected component of the thresholded difference between the source and edited images. We apply pixel and content cues deterministically to the edited image, without re-rendering the underlying edit. We also include SHAM, a zero-dose control that provides the judge-specific sham floor used to contextualize cue-induced shifts. Our pixel-level cues partially overlap with those studied in FRAME and MM-JudgeBias (Hwang et al., 2025; Lee et al., 2026).

3.4 QUALITY-PRESERVATION VERIFICATION

Attributing a cue-induced judgment shift to bias requires establishing that the cue preserves underlying editing quality, i.e., $Q(T(x)) = Q(x)$. Without this condition, the shift could reflect a genuine

change in quality rather than bias. We assess preservation using three complementary sources of evidence: image similarity, calibrated MLLM validators, and human checks.

Similarity. For every cued image, we report whole-image SSIM (Wang et al., 2004) against its un-cued counterpart. For content-site cues, SSIM also captures the added overlay itself and therefore provides only a conservative measure of preservation. We use SSIM as a descriptive measure rather than a filtering criterion.

Calibrated MLLM validators. We use MLLM validators to determine whether a cue changes the underlying edit. To check that they distinguish preserved edits from genuine quality changes, we calibrate them in both directions. The zero-dose SHAM condition measures each validator's false-positive rate, while positive controls measure its sensitivity to known quality changes. For the positive controls, we revert the edit region toward the source image at three strengths and separately blur the edit region. These interventions provide known violations of preservation without requiring human ground truth.

We evaluate three validators on 110 randomly sampled images per cue and 780 SHAM images. For the filtered analysis, we retain only images that pass the two validators with the lower SHAM flag rates. The third validator, from a model family absent from the judge panel, is designed to provide an independent cross-check.

Human checks. We also assess preservation for all ten image-side cues through human annotation, using 30 examples per cue. The first set contains 150 examples covering five pixel-site cues. The second contains 180 examples covering four content-site cues, one pixel-site cue, and 30 SHAM controls; cue identities remain hidden until annotation is complete. Annotators label whether each cue changes the underlying edit as No, Slightly, or Yes. Under the pre-defined criterion, only Yes counts as a preservation failure. Two annotators work independently, with the second annotating a 60-item subset of the blinded set. Appendix C reports the validator pass rates, calibration against positive controls, human-check results, and inter-annotator agreement.

4 Experimental setup

4.1 Evaluation Metrics

Invariance. For each judge and cue, we measure the mean absolute paired change in each 1–10 rating dimension between the un-cued and cued conditions, reported as 10 times the mean absolute paired change. We also report the signed paired mean shift Δ to show the direction of the effect, assess significance with the Wilcoxon signed-rank test, and report bootstrap confidence intervals. To distinguish cue effects from intrinsic variability, we compare the absolute changes against two judge-specific baselines: the change under SHAM and retest noise, estimated by re-evaluating 200 un-cued samples.

For $b \in \{\mathrm{sham}, \mathrm{retest}\}$, let $\widehat{I}_{j,d}^{(b,r)}$ denote the mean absolute paired change (multiplied by 10) in bootstrap resample r. We define each null floor as

$$F_{j,d}^{(b)} = Q_{0.20}\left(\left\{\widehat{I}_{j,d}^{(b,r)}\right\}_{r=1}^{2000}\right), \qquad b \in \{\mathrm{sham}, \mathrm{retest}\}, \tag{2}$$

where Q_p denotes the empirical p-quantile. Let $I_{j,c,d}^{\mathrm{lo}}$ denote the lower end of the corresponding 95% bootstrap interval for the cue-induced absolute change. A cue–dimension cell clears the noise criterion when

$$I_{j,c,d}^{\mathrm{lo}} > \max\left\{F_{j,d}^{\mathrm{sham}}, F_{j,d}^{\mathrm{retest}}\right\}. \tag{3}$$

Agreement. We measure agreement with human judgments using Spearman's correlation between judge scores and human labels, and denote the cue-induced change by $\Delta\rho$. For pairwise evaluation, we measure accuracy against human preferences and test changes with McNemar's test. We cluster bootstrap intervals by editing turn to account for dependence among outputs from the same source and instruction.

Table 1: **Invariance and human agreement across cues and judges.** Cells report 10 times the mean absolute per-item change in each 1–10 rating dimension (IA: instruction adherence; EQ: editing quality; DP: detail preservation). Bold indicates cells whose 95% bootstrap lower bound exceeds both the SHAM and retest floors; plain cells exceed one floor and gray cells exceed neither. Superscripts indicate significant changes in agreement with human ratings (▼ decrease; ▲ increase).

	gpt-5.5			gemini-3.5			kimi-k2.5			qwen3.5			VIEScore		
Bias (short name)	IA	EQ	DP	IA	EQ	DP	IA	EQ	DP	IA	EQ	DP	IA	EQ	DP
Bandwagon	8.2	**11.5**	9.0▼	**18.5**▼	**18.1**	**18.5**	**5.2**	**6.0**	**4.7**	**3.5**	**6.0**	**5.7**	**7.2**	**7.3**▼	**5.0**▼
Authority (model name)	6.4	7.9	6.6	17.8	15.1	17.3	2.9	2.9	3.3	3.2	**4.7**	4.5	4.4	4.9	4.2
Colorfulness (saturation)	5.8	7.3	6.0	8.2▼	7.5	7.6	3.0	2.7	3.2	2.4	3.9	4.3	4.2	4.7	3.8
Aesthetic (aesthetic filter)	6.6▼	8.5▼	7.8	12.0	10.8	12.2	**3.8**	**3.8**	**4.6**	**4.1**	**5.3**	**5.4**	**6.1**	**6.0**	**5.3**
Provenance (watermark)	6.8	7.9▼	7.4▼	7.7▼	8.0	7.6	3.1	**3.3**	**3.6**	**4.6**	**6.9**	**6.2**	**7.1**	**8.2**▼	**6.2**
Luminance (brightness)	6.9	8.5	7.2	7.7	7.5	7.8	**3.7**	**4.1**	**5.1**	**3.9**	**5.7**	**6.1**	5.1	**6.0**	**5.0**
Framing (padding)	7.3	9.5▼	**13.5**▼	8.0	8.1▼	9.3	**4.6**	**4.3**	**7.1**▼	**5.2**	**7.8**	**9.2**	**7.8**	**7.7**	**6.7**
Typographic (text overlay)	9.8	**12.7**	**11.4**	8.3	9.5	9.8	**7.1**	**6.6**	**5.8**	**7.4**	**10.8**	**8.2**	**8.9**	**9.5**▼	**6.5**▼
Verbosity (caption)	6.3	7.6	7.4	17.8▼	15.1	17.5	**4.1**	**4.0**	**4.3**	**5.9**▲	**7.7**▼	**6.4**	**9.3**	**8.7**	**6.6**▼
Attention guidance (box)	9.4	**13.6**▼	**13.6**▼	14.9▼	**16.5**▼	15.1▼	**4.6**	**6.2**	**6.4**	**6.5**	**9.0**	**8.6**	**11.7**▼	**10.9**	**7.5**▼
Scrutiny (inset)	**11.4**	**16.2**▼	**18.1**▼	14.9▼	**21.7**▼	**21.5**▼	**8.3**▼	**9.2**▼	**10.1**▼	**10.2**	**13.7**	**12.5**▼	**12.2**▼	**10.8**▼	**8.8**▼
Distraction (sticker)	**10.8**	**13.5**	**19.5**▼	14.8▼	15.5▼	17.6▼	**10.5**	**9.1**	**8.9**▼	**11.3**	**13.7**	**16.4**	**10.3**	**10.8**▼	**8.1**
sham floor	*6.2*	*7.7*	*6.8*	*17.5*	*15.2*	*16.9*	*2.7*	*2.7*	*3.1*	*2.4*	*3.5*	*3.7*	*4.4*	*4.9*	*3.7*
retest floor	*9.2*	*10.3*	*9.1*	*13.7*	*13.0*	*15.1*	*2.0*	*1.8*	*2.0*	*1.9*	*2.5*	*3.4*	*2.9*	*3.8*	*2.9*

For anchor benchmark a and rating dimension d, let $H_{d,a}$ denote the human labels and let $X_{c,a}$ and $X_{c,a}^{\mathrm{ref}}$ denote the cued and paired reference arms. The cue-induced agreement change is

$$\Delta\rho_{j,c,d,a} = \rho_{\mathrm{S}}(J_{j,d}(X_{c,a}), H_{d,a}) - \rho_{\mathrm{S}}\big(J_{j,d}(X_{c,a}^{\mathrm{ref}}), H_{d,a}\big)\,. \tag{4}$$

Stability. Following CALM (Ye et al., 2025a), we measure pairwise stability using the robustness rate, which captures whether preferences are preserved after swapping the A/B presentation order, and the consistency rate, which measures agreement across repeated evaluations. We also report the slot-preference rate to quantify systematic preference for a presentation position. Let $m(\mathrm{A}) = \mathrm{B}$, $m(\mathrm{B}) = \mathrm{A}$, and $m(\mathrm{Tie}) = \mathrm{Tie}$. Pairwise stability under an A/B swap is

$$S_j = \tfrac{1}{N}\textstyle\sum_{i=1}^{N} \mathbf{1}\left[J_j(x_{i,2}, x_{i,1}) = m(J_j(x_{i,1}, x_{i,2}))\right]. \tag{5}$$

We apply Benjamini–Hochberg correction within each pre-defined family of comparisons. We exclude the SHAM control from the invariance family and treat comparisons for the replication judge in Section 4.2 as a separate family.

4.2 Judges and Evaluation Setup

We audit five judges in three groups: two closed frontier models, gpt-5.5 and gemini-3.5-flash; two models from open-weight families, qwen3.5-plus and kimi-k2.5; and one task-specific judge, VIEScore (Ku et al., 2024a), whose rubric we run on GPT-4o. We request temperature 0 whenever the parameter is configurable; for gpt-5.5, the provider fixes this setting. Retest noise is estimated separately for each judge. We also evaluate a sixth, open-weight judge, qwen3-vl-32b-instruct, on the same samples and cue conditions as a pre-declared replication. We analyze its results as a separate family of comparisons and do not pool them with those of the five primary judges. Appendix B lists the model strings and temperature settings.

5 Results and Discussion

5.1 Invariance under Quality-Preserving Cues

Scoring. Quality-preserving cues can shift ratings beyond judge-specific response noise. Excluding INSET‡, half of the instruction-adherence cells in Table 1 clear both the SHAM and retest floors; the corresponding counts are 34 of 55 for editing quality and 32 of 55 for detail preservation. The effects are not uniform across cues: BANDWAGON clears both floors for most judge–dimension combinations, whereas SATURATION never does. Despite covering at most 4% of the frame, the sticker produces larger rating changes than any pixel-site cue in every judge–dimension cell.

The noise-floor criterion and the signed-shift test capture different properties of the response. A cell that does not clear both noise floors can still exhibit a significant signed shift: the former asks whether the magnitude of change exceeds response variability, whereas the latter tests whether paired changes show a consistent direction. A gray cell in Table 1 should therefore not be read as evidence of no directional effect.

Pairwise. Several cues also shift pairwise preferences, showing that cue sensitivity is not specific to scalar scoring. Figure 4a reports cue-induced changes in pairwise win rates. BANDWAGON increases the cued candidate's win rate under every judge, while the sticker and text overlay tend to disadvantage. By contrast, saturation, the aesthetic filter, and the watermark have little effect on pairwise preferences. The effect of a quality-preserving cue therefore depends not only on the judge and cue, but on whether evaluation is elicited as an absolute score or a comparative preference.

5.2 Agreement with Human Judgments

Table 1 marks significant cue-induced changes in judge–human agreement, which are overwhelmingly negative. Excluding the borderline INSET, 33 of the 34 significant changes are decreases. These decreases occur in 16 of 45 content-site comparisons, compared with 13 of 90 pixel-site and 4 of 30 prompt-site comparisons. This concentration among edit-region cues is descriptive rather than a formal comparison across sites, because the pixel and content groups each include comparisons against both SHAM and the original uncued images.

The pairwise results provide a more direct example. When BANDWAGON assigns a fabricated majority to the candidate rejected by humans, agreement falls for every judge; reversing the fabricated majority instead increases agreement for all five. The same judge can appear more or less aligned with human preferences depending on a quality-irrelevant change in the evaluation context, illustrating why agreement alone is insufficient as a robustness measure.

5.3 Pairwise Stability under Display-Order Swaps

Figure 4b compares verdict-reversal rates under display-order swaps and identical re-queries, showing substantially higher rates under order swaps. This separation holds across the primary judge panel and is especially pronounced for qwen3.5, for which 60.9% of order-swapped pairs reverse compared with 10.5% under identical re-query. The replication judge shows the same qualitative separation between re-query variability and sensitivity to display order.

The two-order rule used by the MT-Bench (Zheng et al., 2023) filters rather than resolves this instability: a winner is retained only when both presentation orders agree. This removes order-sensitive pairs at the cost of coverage, most sharply for qwen3.5. A lenient variant additionally retains the non-Tie choice when only one order returns Tie; pairs whose two orders select opposite winners remain unresolved. The Appendices E and F report the reconciliation analysis and the replication results, respectively.

5.4 Preservation Evidence

Our preservation checks support treating most image-side cues as quality-preserving at the cue level. Figure 4c shows the validators' responses to controlled degradations. Among the ten image-side cues, only INSET shows a significant departure from paired SHAM images. In the sham-controlled human annotation set, none of the five evaluated cues differs significantly from SHAM under the prespecified criterion.

The evidence for INSET is less clear. Two validators flag it more often than SHAM, and a more inclusive post-hoc human criterion also identifies a difference. These signals do not justify treating the cue as uniformly quality-preserving, but neither do they overturn the preservation evidence for the remaining image-side interventions. We therefore treat INSET as borderline, report its results separately, and exclude it from aggregate counts. This separation prevents the ambiguous cue from affecting the main aggregate conclusions while retaining its individual results for transparency.

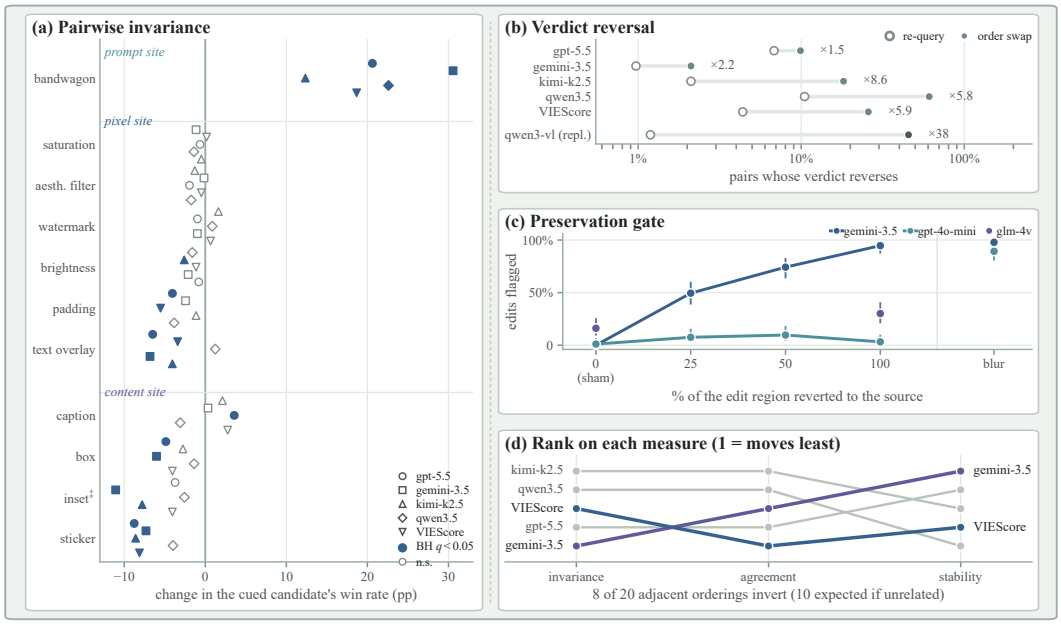

Figure 4: **Main results of the EditJudgeBias audit.** (a) Pairwise preference shifts induced by quality-preserving cues. (b) Verdict reversal rates under display-order swaps, compared with identical re-queries. (c) Calibration of the preservation validators using controlled edit degradation and the zero-dose control. (d) Comparison of judges across invariance, agreement, and stability, showing that judge rankings differ across the three robustness measures.

5.5 Robustness across Measures

Figure 4d shows that the three robustness measures induce different judge orderings. kimi-k2.5 ranks first on invariance and agreement but third on stability, whereas gemini-3.5 moves from fifth on invariance to first on stability. Across adjacent measure pairs, 8 of 20 judge orderings reverse, compared with 10 under the unrelated-measure reference. These differences indicate that invariance, agreement, and stability capture distinct aspects of judge behavior rather than interchangeable estimates of a single robustness quantity. A single aggregate score would therefore obscure which failure mode is responsible for a judge's apparent weakness. Appendix D reports the exact rankings and alternative summaries.

6 Conclusion

We introduced **EditJudgeBias** to test how image-editing judges respond to controlled cues that are designed to preserve the underlying edit. Across the five primary judges, such cues can shift both ratings and pairwise preferences beyond ordinary response variability. Human agreement does not fully predict this behavior: display-order swaps can reverse preferences, and the same judges rank differently on invariance, agreement, and stability. More broadly, our findings suggest that judge reliability should not be defined solely by whether a judge agrees with humans, but also by whether its decisions remain invariant when the underlying quality being evaluated is unchanged. This offers an insight for MLLM-based evaluation: reliable judging requires measuring the intended construct rather than responding to incidental features of how that construct is presented.

AI DISCLOSURE

Generative AI tools assisted with drafting and language editing, figure preparation, and software tasks including code review and repository maintenance. The research questions, literature selection, experimental design and deployment, data analysis, and interpretation were carried out by the authors.

ETHICS STATEMENT

EditJudgeBias uses existing published image-editing benchmarks and controlled interventions intended to audit automated evaluators. Image-side cues are checked for preservation, and human annotation is used only for this purpose. Some prompt-side cues deliberately present fabricated social information, such as a claimed majority preference, to test susceptibility to information irrelevant to editing quality. These statements are confined to the evaluation protocol and are not treated as factual benchmark annotations.

REPRODUCIBILITY STATEMENT

We provide an anonymized repository at reproducibility repository. This release contains code, configurations, prompts, and scripts, but excludes benchmark images, human annotation records, and experimental judge/validator responses. The author-generated I2EBench outputs used in our experiments are also not included. The release therefore does not by itself permit exact reconstruction of all evaluated inputs or reported results.

Several judges are accessed through external APIs, so repeated identical calls may produce different outputs. Our design explicitly accounts for this variability using judge-specific re-query baselines, SHAM controls, paired inference, confidence intervals, multiple-testing correction, and a separately analyzed replication arm. Reproducibility therefore concerns the statistical findings and qualitative conclusions under expected API variability rather than exact reproduction of every model response.

REFERENCES

Amit Agarwal, Hitesh Laxmichand Patel, Meizhu Liu, Jyotika Singh, Karan Dua, Hansa Meghwani, Matthew Rowe, Michael Avendi, Yassi Abbasi, Tao Sheng, et al. Do image–text metrics respect semantic invariances? In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 39089–39116, 2026.

Xuehai Bai, Yang Shi, Yi-Fan Zhang, Xuanyu Zhu, Yuran Wang, Yifan Dai, Xinyu Liu, Yiyan Ji, Xiaoling Gu, and Yuanxing Zhang. Edit-compass & editreward-compass: A unified benchmark for image editing and reward modeling. *arXiv preprint arXiv:2605.13062*, 2026.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18392–18402. IEEE, 2023.

Honghua Chen, Zitong Xu, Huiyu Duan, Xinyun Zhang, Xiongkuo Min, and Guangtao Zhai. Reasonedit: Towards interpretable image editing evaluation via reinforcement learning. *arXiv preprint arXiv:2605.07477*, 2026a.

Jianlin Chen, Wenhui Chen, Ziyao Lin, and Chi Man Vong. A judge should know what changed: Construct validity for llm-as-a-judge evaluation. *arXiv preprint arXiv:2608.24419*, 2026b.

Tianyu Chen, Yasi Zhang, Zhi Zhang, Peiyu Yu, Shu Wang, Zhendong Wang, Kevin Lin, Xiaofei Wang, Zhengyuan Yang, Linjie Li, Chung-Ching Lin, Jianwen Xie, Oscar Leong, Lijuan Wang, Yingnian Wu, and Mingyuan Zhou. EdiVal-Agent: An object-centric framework for automated, fine-grained evaluation of multi-turn editing. In *International Conference on Learning Representations*, volume 2026, pp. 114919–114965, 2026c.

Zhaorun Chen, Zichen Wen, Yichao Du, Yiyang Zhou, Chenhang Cui, Siwei Han, Jen Weng, Chaoqi Wang, Zhengwei Tong, Leria Huang, Canyu Chen, Haoqin Tu, Qinghao Ye, Zhihong Zhu, Yuqing Zhang, Jiawei Zhou, Zhuokai Zhao, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. MJ-Bench:

Is your multimodal reward model really a good judge for text-to-image generation? In *Advances in Neural Information Processing Systems 38*, pp. 69174–69228. Neural Information Processing Systems Foundation, 2025.

Hao Cheng, Erjia Xiao, Jindong Gu, Le Yang, Jinhao Duan, Jize Zhang, Jiahang Cao, Kaidi Xu, and Renjing Xu. Unveiling typographic deceptions: Insights of the typographic vulnerability in large vision-language models. In *European Conference on Computer Vision*, pp. 179–196. Springer, 2024.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL https://arxiv.org/abs/2505.14683.

Haiwen Diao, Jiahao Wang, Chenjing Ding, Hanming Deng, Jiangnan Chen, Ruixi Zhang, Ruohui Wang, Wenwen Tong, Xiangyu Fan, Yubo Wang, Yue Zhu, Yuwei Niu, Zhengqi Bai, Zhiqian Lin, Zhitao Yang, Zhongang Cai, Bo Yang, Chen Feng, Chengguang Lv, Guangjia Liu, Guanlin Wang, Hanyu Zhang, Haojia Yu, Hongcan Xiao, Hongli Wang, Huan Wu, Huaping Zhong, Jian Fang, Jianan Fan, Jiaqi Li, Jiefan Lu, Jing Zuo, Jingcheng Ni, Junxiang Xu, Linjun Dai, Mutian Xu, Peishen Yan, Penghao Wu, Ruijie Mao, Ruisi Wang, Shihao Bai, Shuang Yang, Shuya Yang, Shuyan Zheng, Silei Wu, Siying Li, Tao Chu, Tianbo Zhong, Tongxi Zhou, Weichao Luo, Weichen Fan, Wenhao Jia, Wenjie Gao, Xiangli Kong, Yan Li, Yang Yong, Zimo Wen, Zixuan Qian, Wenxiu Sun, Ruihao Gong, Quan Wang, Lewei Lu, Lei Yang, Ziwei Liu, and Dahua Lin. Sensenova-u1.5: Towards native unified visual intelligence, 2026. URL https://arxiv.org/abs/2609.11929.

Fengyi Fu, Mengqi Huang, Shaojin Wu, Yunsheng Jiang, Yufei Huo, Hao Li, Yinghang Song, Fei Ding, Jianzhu Guo, Qian He, Zheren Fu, Zhendong Mao, and Yongdong Zhang. Lance: Unified multimodal modeling by multi-task synergy, 2026. URL https://arxiv.org/abs/2605.18678.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International conference on machine learning*, pp. 10835–10866. PMLR, 2023.

Shiqi Gao, Zitong Xu, Kang Fu, Huiyu Duan, Xiongkuo Min, and Jia Wang. Evaluating image editing with llms: A comprehensive benchmark and intermediate-layer probing approach. *Displays*, pp. 103494, 2026.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Yichen Huang and Timothy Baldwin. Robustness tests for automatic machine translation metrics with adversarial attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 5126–5135, 2023.

Yerin Hwang, Dongryeol Lee, Kyungmin Min, Taegwan Kang, Yongil Kim, and Kyomin Jung. Fooling the LVLM judges: Visual biases in LVLM-based evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23186–23205, Suzhou, China, November 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.emnlp-main.1182/.

Martina Ianaro, Guilherme Fernandes, Maurizio Gabbrielli, and Joao Magalhaes. Order matters: Lvlms as judges for temporal reasoning in image sequences. *arXiv preprint arXiv:2608.10908*, 2026.

Fengxian Ji, Jingpu Yang, Zirui Song, Lang Gao, Junhong Liang, Zhenhao Chen, Jinghui Zhang, and Xiuying Chen. ServImage: An image generation and editing benchmark from real-world

commercial imaging services. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43504–43529. Association for Computational Linguistics, 2026.

Dongfu Jiang, Max Ku, Tianle Li, Yuansheng Ni, Shizhuo Sun, Rongqi Fan, and Wenhu Chen. Genai arena: An open evaluation platform for generative models. *Advances in Neural Information Processing Systems*, 37:79889–79908, 2024.

Zhangqi Jiang, Zheng Sun, Xianfang Zeng, Yufeng Yang, Xuanyang Zhang, Yongliang Wu, Wei Cheng, Gang Yu, Xu Yang, and Bihan Wen. Geditbench v2: A human-aligned benchmark for general image editing. *arXiv preprint arXiv:2603.28547*, 2026.

Shuitsu Koyama, Yuiga Wada, Daichi Yashima, and Komei Sugiura. Mllm-as-a-judge exhibits model preference bias. *arXiv preprint arXiv:2604.11589*, 2026.

Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhu Chen. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12268–12290, 2024a.

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *International Conference on Learning Representations*, volume 2024, pp. 46689–46722, 2024b.

Seongyun Lee, Seungone Kim, Sue 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*, pp. 11286–11315, 2024.

Sua Lee, Sanghee Park, and Jinbae Im. Mm-judgebias: A benchmark for evaluating compositional biases in mllm-as-a-judge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25336–25373, 2026.

Runzhou Liu, Hailey Weingord, Sejal Mittal, Prakhar Dungarwal, Anusha Nandula, Bo Ni, Samyadeep Basu, Hongjie Chen, Nesreen K Ahmed, Li Li, et al. Human-aligned mllm judges for fine-grained image editing evaluation: A benchmark, framework, and analysis. *arXiv preprint arXiv:2602.13028*, 2026.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025a.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. Rm-bench: Benchmarking reward models of language models with subtlety and style. In *International Conference on Learning Representations*, volume 2025, pp. 44323–44355, 2025b.

Xin Luo, Jiahao Wang, Chenyuan Wu, Shitao Xiao, Xiyan Jiang, Defu Lian, Jiajun Zhang, Dong Liu, and Zheng Liu. Editscore: Unlocking online rl for image editing via high-fidelity reward modeling. In *International Conference on Learning Representations*, volume 2026, pp. 33027–33056, 2026.

Yiwei Ma, Jiayi Ji, Ke Ye, Weihuang Lin, Zhibin Wang, Yonghan Zheng, Qiang Zhou, Xiaoshuai Sun, and Rongrong Ji. I2ebench: A comprehensive benchmark for instruction-based image editing. *Advances in Neural Information Processing Systems*, 37:41494–41516, 2024.

Justin D Norman, Michael U Rivera, and D Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of llm-as-a-judge models across agreement, consistency, and bias. *arXiv preprint arXiv:2606.19544*, 2026.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 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. *arXiv preprint arXiv:2606.02578*, 2026.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4902–4912, 2020.

Subhadeep Roy, Gagan Bhatia, and Steffen Eger. Prototypicality bias reveals blindspots in multimodal evaluation metrics. *arXiv preprint arXiv:2601.04946*, 2026.

Ananya B Sai, Tanay Dixit, Dev Yashpal Sheth, Sreyas Mohan, and Mitesh M Khapra. Perturbation checklists for evaluating nlg evaluation metrics. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7219–7234, 2021.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*, 2023.

Ningyu Sun, Zhaolin Cai, Zitong Xu, Peihang Chen, Huiyu Duan, Yichao Yan, Xiongkuo Min, and Xiaokang Yang. Fine-grained human pose editing assessment via layer-selective mllms. *arXiv preprint arXiv:2601.10369*, 2026.

Doan Nam Long Vu, Nafise Sadat Moosavi, and Steffen Eger. Layer or representation space: What makes bert-based evaluation metrics robust? In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 3401–3411, 2022.

Jifang Wang, Xue Yang, Longyue Wang, Zhenran Xu, Yiyu Wang, Yaowei Wang, Weihua Luo, Kaifu Zhang, Baotian Hu, and Min Zhang. A unified agentic framework for evaluating conditional image generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12626–12646, 2025a.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9440–9450, 2024.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zihan Wang, Guansong Pang, Zelin Liu, Wenjun Miao, Jin Zheng, and Xiao Bai. On the adversarial robustness of multimodal llm judges. *arXiv preprint arXiv:2606.15608*, 2026.

Keming Wu, Sicong Jiang, Max Ku, Ping Nie, Minghao Liu, and Wenhu Chen. EditReward: A human-aligned reward model for instruction-guided image editing. In *International Conference on Learning Representations*, volume 2026, pp. 46591–46622, 2026.

Yongliang Wu, Zonghui Li, Xinting Hu, Xinyu Ye, Xianfang Zeng, Gang Yu, Wenbo Zhu, Bernt Schiele, Ming-Hsuan Yang, and Xu Yang. KRIS-Bench: Benchmarking next-level intelligent image editing models. In *Advances in Neural Information Processing Systems 38*, pp. 174202–174248. Neural Information Processing Systems Foundation, 2025a.

Zhaofeng Wu, Michihiro Yasunaga, Andrew Cohen, Yoon Kim, Asli Celikyilmaz, and Marjan Ghazvininejad. rewordbench: Benchmarking and improving the robustness of reward models with transformed inputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3383–3409, 2025b.

Jinheng Xie, Zhenheng Yang, and Mike Zheng Shou. Show-o2: Improved native unified multimodal models. In *Advances in Neural Information Processing Systems 38*, pp. 53002–53030. Neural Information Processing Systems Foundation, 2025.

Zitong Xu, Huiyu Duan, Bingnan Liu, Guangji Ma, Jiarui Wang, Liu Yang, Shiqi Gao, Xiaoyu Wang, Jia Wang, Xiongkuo Min, et al. Lmm4edit: Benchmarking and evaluating multimodal image editing with lmms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6908–6917, 2025.

Zitong Xu, Huiyu Duan, Xinyun Zhang, Weifei Xiong, Tianyi Zheng, Xiongkuo Min, Qiang Hu, Zhengxue Cheng, Bo Li, and Guangtao Zhai. Miescore: Human-aligned evaluation for multi-source image editing. *arXiv preprint arXiv:2608.02059*, 2026.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. Justice or prejudice? quantifying biases in llm-as-a-judge. In *International Conference on Learning Representations*, volume 2025, pp. 102351–102390, 2025a.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. ImgEdit: A unified image editing dataset and benchmark. In *Advances in Neural Information Processing Systems 38*, pp. 146002–146032. Neural Information Processing Systems Foundation, 2025b.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in neural information processing systems*, 36:31428–31449, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, Hua Yang, Xue Yang, and Haodong Duan. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. In *Advances in Neural Information Processing Systems 38*, pp. 192865–192904. Neural Information Processing Systems Foundation, 2025.

Xiangyu Zhao, Peiyuan Zhang, Junming Lin, Tianhao Liang, Yuchen Duan, Shengyuan Ding, Changyao Tian, Yuhang Zang, Junchi Yan, and Xue Yang. Trust your critic: Robust reward modeling and reinforcement learning for faithful image editing and generation. *arXiv preprint arXiv:2603.12247*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Xiaohan Zou, Roshan Sridhar, Mohammadtaher Safarzadeh, and Dan Roth. When vision-language models judge without seeing: Exposing informativeness bias. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15417–15448, 2026.

A LIMITATIONS

Our study has three limitations. First, EditJudgeBias evaluates a finite set of controlled cues at specified intervention strengths; interactions among co-occurring cues and responses across a broader range of strengths remain to be characterized. Second, the results reflect the evaluated judge versions and configurations. As models and API services evolve, effect magnitudes and relative judge rankings may change, motivating periodic re-evaluation with the same audit protocol. Third, our experiments focus on diagnosing cue sensitivity, while mitigation is explored only through limited aggregation analyses and a preliminary prompting pilot. Developing and systematically evaluating mitigation strategies under the main experimental protocol remains future work. Our preservation checks focus on the requested edit; rubric-aligned human ratings of both versions are needed to distinguish evaluator bias from legitimate visual-quality or detail-preservation penalties.

B BENCHMARK AND EVALUATION PROTOCOL

B.1 EDITING SAMPLES AND POOLS

The judged subset contains 1,196 samples drawn from five published editing benchmarks. These benchmarks represent three independent content pools rather than five: the rated ImagenHub images are a subset of the MagicBrush development set, and most GenAI-Bench editing prompts come from the same pool. Figure 2b marks the benchmarks that share underlying content.

The subset is divided into one *breadth* block and two *anchor* blocks. The breadth block contains 611 samples stratified by source and edit type and is used for the invariance analysis. The anchor blocks retain complete editing turns, keeping multiple editors' outputs for each source–instruction pair together. EBench-18K contributes 48 turns with 8 of the block's 17 editors, for 384 samples; ImagenHub contributes 30 turns with 8 editors, for 240 samples. These blocks provide the human labels used in the agreement analysis and are also the blocks on which we re-score editor leaderboards. Across these blocks there are 1,235 sample assignments; 39 triplets occur in both the breadth and anchor blocks, yielding 1,196 unique triplets overall. The pairwise protocol uses 616 pairs of edits for the same request, drawn from 1,086 edited images; 438 pairs have a decisive human preference.

B.2 CUES, OPERATORS AND DOSES

All image-side cues are deterministic PIL operations applied after the edit has been produced. For the global pixel-site cues, SATURATION and BRIGHTNESS scale the corresponding enhancement by 1.2. The AESTHETIC FILTER increases contrast by 1.1, shifts the red and blue channels by $+8$ and -8 points for a warm cast, and applies an unsharp mask at 60% with radius 1/256 of the frame. PADDING adds a black border equal to 10% of each image dimension on every side. WATERMARK places "Reference Image" in the bottom-right corner at 45% opacity, using a font size of 4% of the short edge and a 3% inset. TEXT OVERLAY places a salient word from the instruction in a semi-transparent band along the top edge at 50% opacity, with a font size of 5% of the short edge.

The four content-site cues are defined relative to the edit region. BOX draws a red rectangle around the region with a stroke width of 3/512 of the frame. INSET crops and magnifies the region by $2\times$, caps its side length at 40% of the frame, and places it in a corner with a 3/512 white border. CAPTION prints "Edited: {instruction}", truncated to 14 words, in a white bar with near-black text at a font size of 3.5% of the short edge. STICKER places an unrelated element covering at most 4% of the frame area outside the edit region, inset by 2%. The zero-dose SHAM control is a JPEG round trip at quality 95.

The prompt-site cues leave the image unchanged. In scoring, BANDWAGON appends "Note: previous reviewers rated this edit very highly." to the judge prompt, while MODEL NAME adds "This edited image was produced by the {model} model." using the generator label recorded for the edit in its source dataset. In pairwise evaluation, the bandwagon message becomes "Note: 90% of human evaluators preferred Image {A|B}.", with the claimed majority assigned to a seeded side so that it does not systematically favor slot A; the model-name cue names both candidates. The protocol-site cue POSITION simply swaps the display order of the pair, changing neither the image pixels nor the prompt text.

B.3 THE EDIT REGION

The four content-site cues require an estimate of the edit region. We use the benchmark-provided source mask when available. Otherwise, we threshold the per-channel difference between the source and edited images at 15 out of 255 and take the largest connected component, falling back to a centered box if no component is found. BOX and INSET operate directly on this region; STICKER is placed away from it, while CAPTION uses the region only to avoid covering the edit.

B.4 JUDGES, RUBRICS AND COLLECTION

In the scoring protocol, each judge receives the source image, editing instruction, and edited image and returns one integer from 1 to 10 for instruction adherence, editing quality, and detail preservation. Their sum gives the 3–30 score used in the analysis. We request temperature 0 whenever this parameter is configurable; gpt-5.5 uses the provider-fixed setting. Retest noise is estimated by re-querying each judge on 200 un-cued items. VIEScore uses its original rubric with GPT-4o as the underlying model rather than constituting a separate model family.

The API model strings are `gpt-5.5`, `gemini-3.5-flash`, `kimi-k2.5`, and `qwen3.5-plus` for the primary general-purpose judges, and `gpt-4o` for VIEScore. The replication judge uses `qwen3-vl-32b-instruct`; the three preservation validators use `gemini-3.5-flash`, `gpt-4o-mini`, and `glm-4v`. Calls are made through an OpenAI-compatible relay that does not expose provider-side checkpoint identifiers, so these model strings are the most specific version identifiers available to us.

The complete scoring, pairwise, and preservation-validation prompts, together with the prompt-site cue strings and human annotation instructions, are reproduced in Appendix G.

C QUALITY-PRESERVATION VALIDATION

C.1 VALIDATOR CALIBRATION AND RESULTS

The preservation gate is evaluated independently of scoring and pairwise judging. Each validator receives the source image, instruction, edited image, and cued edited image, and checks whether the cue changes instruction adherence, editing quality, detail preservation, or scene semantics; a sample passes only when all four answers are no. We evaluate each validator on 110 images per cue, with 780 SHAM images used to estimate its baseline flag rate. Gemini-3.5 and gpt-4o-mini, which have the lower SHAM flag rates, form the gate, while glm-4v serves as an independent cross-check. Table 2 reports the cue-level pass rates and calibration against controlled degradations.

Table 2: **Preservation validation across cues.** Panel A reports validator pass rates; Panel B reports detection of controlled quality degradations among 93 evaluable examples per condition.

Panel A: Preservation pass rates (%).

Cue	gemini-3.5	gpt-4o-mini	glm-4v
Colorfulness (saturation)	100.0	97.3	94.5
Aesthetic (aesthetic filter)	99.1	95.5	87.3
Provenance (watermark)	100.0	98.2	91.8
Luminance (brightness)	99.1	97.3	93.6
Framing (padding)	100.0	94.5	93.6
Typographic (text overlay)	100.0	90.0	92.7
Verbosity (caption)	100.0	99.1	94.5
Attention guidance (box)	100.0	97.3	89.1
Scrutiny (inset)	91.8	96.4	73.6
Distraction (sticker)	98.2	96.4	90.9
SHAM	99.6	97.6	89.7

Panel B: Controlled-degradation detection.

Validator	25% reversion	50% reversion	100% reversion	Blur
gemini-3.5	46	69	88	91
gpt-4o-mini	7	9	3	83
glm-4v	—	—	28	—

Against paired SHAM images, INSET is the only cue with a significantly lower pass rate after Benjamini–Hochberg correction: the difference is detected by gemini-3.5 ($q = 0.020$) and glm-4v ($q < 0.001$), but not by gpt-4o-mini. A disjoint repeat over seven cues yields the same qualitative pattern. Gemini-3.5 again differs from SHAM only for INSET, with a pass rate of 87.3% versus 100.0% ($q < 0.001$), while gpt-4o-mini shows no significant difference for any repeated cue.

The controlled degradations clarify what each validator detects. Gemini-3.5 becomes increasingly sensitive as more of the edit region is reverted toward the source, whereas gpt-4o-mini detects edit-region blur but rarely flags reversion (Table 2, Panel B). The two validators are therefore not equally informative about instruction-adherence preservation, and requiring both to pass provides limited additional evidence beyond the primary validator on this dimension.

C.2 Human Verification

We also assess preservation through human annotation, with 30 examples for each of the ten image-side cues. The first annotation package covers five pixel-site cues; the second covers four content-site cues and one pixel-site cue together with SHAM controls, with cue identities hidden until annotation is complete. Annotators label each intervention as changing the underlying edit by No, Slightly, or Yes, with only Yes prespecified as a preservation failure.

Under this criterion, none of the five cues in the sham-controlled package differs significantly from SHAM: the inset has 2/30 failures versus 0/30 for SHAM, while the other four cues have none. A post-hoc criterion that also counts Slightly as a failure changes the conclusion only for the inset, with 11/30 failures versus 0/30 for SHAM ($q = 0.002$). A second annotator independently labels 60 items from this package, yielding 98.3% raw agreement and Cohen's $\kappa = 0.893$. The annotators were aware of the study hypothesis.

C.3 Sensitivity Analysis on Validated Images

As a separate sensitivity analysis, we repeat the breadth analysis using only images that pass both gating validators, leaving 47–57 items per cue–judge cell. All 24 effects that are significant in both the full and filtered analyses retain the same direction. We therefore use this analysis only as a check on effect direction within the validated subset, rather than as a second estimate of effect magnitude or evidence of preservation for every image.

Table 4: **Signed shifts in editing quality.** Cells report paired mean changes on the 1–10 scale; **bold** indicates Benjamini–Hochberg significance. ‡ marks the borderline cue.

Cue	gpt-5.5	gemini-3.5	kimi-k2.5	qwen3.5	VIEScore
Bandwagon	**+0.62**	**+0.41**	**+0.50**	**+0.42**	**+0.48**
Authority (model name)	**−0.13**	−0.05	**+0.10**	**+0.18**	+0.01
Colorfulness (saturation)	**−0.10**	**−0.17**	**−0.07**	**−0.08**	+0.02
Aesthetic (aesthetic filter)	−0.09	**−0.18**	**−0.12**	−0.02	**−0.09**
Provenance (watermark)	**−0.32**	**−0.28**	+0.02	**−0.38**	**−0.25**
Luminance (brightness)	**−0.22**	**−0.25**	**−0.22**	**−0.25**	**−0.17**
Framing (padding)	**−0.45**	**−0.34**	+0.02	**−0.44**	**−0.31**
Typographic (text overlay)	**−1.07**	**−0.77**	**−0.58**	**−0.97**	**−0.55**
Verbosity (caption)	**−0.24**	−0.16	**−0.16**	**−0.16**	**+0.11**
Attention guidance (box)	**−0.95**	**−1.17**	**−0.55**	**−0.64**	**−0.25**
Scrutiny (inset)‡	**−1.17**	**−2.15**	**−0.81**	**−1.03**	**−0.15**
Distraction (sticker)	**−1.06**	**−1.41**	**−0.88**	**−1.26**	**−0.82**

D Detailed Judge Results and Robustness Analyses

D.1 Detailed Scoring Results

Table 1 reports the magnitude of cue-induced rating changes irrespective of direction. Tables 3, 4, and 5 complement this view with the paired mean signed shifts on the original 1–10 scale for instruction adherence, editing quality, and detail preservation, respectively. Changes in agreement with human ratings are marked separately in Table 1.

Unparsed responses are excluded, leaving 597–611 paired observations per judge for the scoring analyses and 361–384 for the agreement analyses. The reference arm follows the collection design: bandwagon, authority, colorfulness, aesthetic, provenance, verbosity, scrutiny, and distraction are compared with SHAM, whereas luminance, framing, typographic, and attention guidance are compared with the original un-cued arm. Because the pixel and content sites each mix these reference conditions, the site-level rates reported in the main text are descriptive rather than formal comparisons across sites.

Table 3: **Signed shifts in instruction adherence.** Cells report paired mean changes on the 1–10 scale; **bold** indicates Benjamini–Hochberg significance. ‡ marks the borderline cue.

Cue	gpt-5.5	gemini-3.5	kimi-k2.5	qwen3.5	VIEScore
Bandwagon	**+0.45**	**+0.44**	**+0.34**	**+0.24**	**+0.45**
Authority (model name)	−0.04	**+0.12**	+0.05	**+0.13**	**−0.10**
Colorfulness (saturation)	+0.00	−0.09	+0.08	+0.01	+0.10
Aesthetic (aesthetic filter)	+0.10	+0.05	+0.06	+0.09	**+0.12**
Provenance (watermark)	**−0.20**	**−0.19**	+0.02	**−0.29**	+0.01
Luminance (brightness)	−0.03	**−0.11**	−0.03	−0.02	+0.08
Framing (padding)	**−0.25**	**−0.16**	+0.05	**−0.18**	**−0.26**
Typographic (text overlay)	**−0.67**	**−0.40**	**−0.59**	**−0.56**	−0.17
Verbosity (caption)	−0.00	**+0.19**	+0.03	+0.14	**+0.39**
Attention guidance (box)	**−0.39**	**−0.35**	**−0.26**	**−0.34**	+0.03
Scrutiny (inset)‡	**−0.52**	**−0.85**	**−0.57**	**−0.60**	**+0.27**
Distraction (sticker)	**−0.76**	**−0.71**	**−0.99**	**−1.04**	**−0.63**

D.2 Robustness across Aggregation Schemes

Table 6 gives the exact judge rankings underlying Figure 4d and shows how the observed rank inversions change under alternative summary schemes.

Table 5: **Signed shifts in detail preservation.** Cells report paired mean changes on the 1–10 scale; **bold** indicates Benjamini–Hochberg significance. ‡ marks the borderline cue.

Cue	gpt-5.5	gemini-3.5	kimi-k2.5	qwen3.5	VIEScore
Bandwagon	**+0.51**	**+0.69**	**+0.22**	**+0.21**	**+0.30**
Authority (model name)	+0.02	**+0.46**	−0.01	+0.03	**−0.12**
Colorfulness (saturation)	−0.06	+0.03	−0.05	−0.08	+0.02
Aesthetic (aesthetic filter)	**−0.17**	**+0.16**	**−0.24**	**−0.10**	**−0.14**
Provenance (watermark)	**−0.36**	−0.04	+0.02	**−0.25**	**+0.19**
Luminance (brightness)	**−0.30**	−0.09	**−0.32**	**−0.25**	**−0.16**
Framing (padding)	**−1.06**	**−0.42**	**−0.45**	**−0.61**	**−0.40**
Typographic (text overlay)	**−0.92**	**−0.57**	**−0.31**	**−0.58**	**−0.21**
Verbosity (caption)	−0.07	**+0.46**	+0.03	**+0.11**	**+0.20**
Attention guidance (box)	**−0.86**	**−0.42**	**−0.31**	**−0.36**	**−0.11**
Scrutiny (inset)‡	**−1.32**	**−1.98**	**−0.75**	**−0.83**	−0.02
Distraction (sticker)	**−1.71**	**−1.46**	**−0.74**	**−1.43**	**−0.45**

Table 6: **Robustness rankings across judges.** Panel A gives the primary rankings used in Figure 4d; Panel B summarizes rank inversions under alternative aggregation schemes.

Panel A: Primary robustness rankings.

Judge	Invariance	Agreement	Stability
gpt-5.5	9.67 (4)	0.0474 (4)	0.8458 (2)
gemini-3.5	12.94 (5)	0.0402 (3)	0.8994 (1)
kimi-k2.5	5.29 (1)	0.0272 (1)	0.7955 (3)
qwen3.5	7.14 (2)	0.0348 (2)	0.3523 (5)
VIEScore	7.32 (3)	0.0615 (5)	0.6883 (4)

Panel B: Rank inversions under alternative summaries.

Summary scheme	Adjacent-pair inversions	Unrelated reference
Mean cell	**8/20**	10
Worst cell	**9/20**	10
Significance-based	**6/20** (6 ties)	7.0 over 14 untied pairs

For the primary summary, invariance is the mean of the mean absolute paired changes over the 36 non-sham cue–dimension cells, agreement is the mean absolute change in Spearman's ρ over anchor cells, and stability is RR.. The worst-cell summary replaces the first two means by their maxima. The significance-based summary uses counts of significant invariance and agreement cells and CR for stability.

E Additional Analyses

E.1 Control and Cross-Benchmark Analyses

The following analyses test whether the main effects can be attributed to narrower features of the evaluation set. If the photometric results arise mainly because a cue repeats or conflicts with the requested edit, they should weaken once such instruction–cue collisions are removed (Table 7). Likewise, effects driven primarily by poor edits should attenuate among outputs with the highest human mean-opinion scores (Table 8). We also examine the four cues shared by both anchor benchmarks to distinguish benchmark-specific changes in human agreement from patterns that recur across anchors (Table 9).

Table 7: **Photometric controls for instruction collisions.** Cells report paired mean score shifts for clean and collision items with 95% confidence intervals; **bold** indicates Benjamini–Hochberg significance.

		clean items	collision items
Cue	Judge	Δ 95% CI	Δ 95% CI
Luminance (brightness)	gpt-5.5	**−0.67** [−0.89, −0.45]	−0.06 [−0.65, +0.52]
	gemini-3.5	−0.16 [−0.35, +0.04]	**−1.65** [−2.49, −0.82]
	kimi-k2.5	**−0.69** [−0.86, −0.52]	−0.12 [−0.43, +0.20]
	qwen3.5	**−0.66** [−0.95, −0.39]	+0.07 [−0.72, +0.92]
	VIEScore	**−0.33** [−0.54, −0.12]	+0.02 [−0.41, +0.47]
Colorfulness (saturation)	gpt-5.5	−0.18 [−0.37, +0.03]	−0.10 [−0.56, +0.34]
	gemini-3.5	−0.00 [−0.19, +0.20]	**−1.12** [−1.98, −0.27]
	kimi-k2.5	−0.08 [−0.24, +0.08]	+0.13 [−0.10, +0.37]
	qwen3.5	**−0.18** [−0.41, +0.06]	−0.01 [−0.59, +0.49]
	VIEScore	+0.18 [−0.00, +0.37]	−0.01 [−0.39, +0.35]
Aesthetic (aesthetic filter)	gpt-5.5	**−0.37** [−0.61, −0.13]	+0.65 [+0.13, +1.18]
	gemini-3.5	+0.37 [+0.06, +0.70]	**−1.37** [−2.22, −0.53]
	kimi-k2.5	**−0.43** [−0.58, −0.28]	+0.17 [−0.14, +0.50]
	qwen3.5	−0.14 [−0.41, +0.15]	+0.40 [−0.31, +1.15]
	VIEScore	**−0.23** [−0.46, −0.01]	+0.40 [−0.06, +0.89]

Table 8: **Cue effects on high-quality edits.** Cells report mean signed score shifts within the highest human-MOS tertile; **bold** indicates Benjamini–Hochberg significance.

Cue	gpt-5.5	gemini-3.5	kimi-k2.5	qwen3.5	VIEScore
Bandwagon	**+2.65**	**+3.89**	**+1.36**	**+0.74**	**+1.30**
Luminance (brightness)	−0.28	**+2.67**	**−0.59**	−1.28	−0.57
Framing (padding)	**−2.83**	**+1.48**	**−1.17**	**−1.59**	**−0.89**
Typographic (text overlay)	**−3.28**	−0.30	**−1.87**	**−3.61**	**−1.48**
Attention guidance (box)	**−4.22**	−1.09	**−2.33**	**−2.91**	−2.33
Distraction (sticker)	**−4.20**	**−2.26**	**−4.13**	**−5.69**	**−2.50**

The controls do not support either explanation as sufficient for the strongest effects. For luminance, significant negative shifts remain on clean items for four of the five judges, so the effect is not confined to photometric instruction collisions. Restricting the analysis to the highest human-MOS tertile likewise leaves several large effects intact: BANDWAGON remains positive and the sticker negative for all five judges.

Table 9: **Cross-benchmark agreement and selection-corrected inference.** Panel A reports $\Delta\rho$ for the four cues evaluated on both anchor benchmarks; **bold** indicates that the uncorrected turn-clustered 95% confidence interval excludes zero. Panel B reports permutation tests with p_{sel} correcting for selection of the most extreme cue.

Panel A: Changes in judge–human rank correlation.

Cue	Anchor	gpt-5.5	gemini-3.5	kimi-k2.5	qwen3.5	VIEScore
Luminance (brightness)	EBench-18K	−0.006	−0.008	−0.010	+0.013	−0.001
	ImagenHub	+0.060	**−0.049**	−0.028	−0.035	+0.019
Framing (padding)	EBench-18K	**−0.105**	**−0.044**	−0.004	−0.018	−0.002
	ImagenHub	+0.014	−0.018	−0.045	−0.017	+0.014
Typographic (text overlay)	EBench-18K	−0.025	−0.007	+0.008	+0.012	−0.037
	ImagenHub	+0.011	**−0.055**	−0.037	**−0.066**	**−0.066**
Attention guidance (box)	EBench-18K	**−0.073**	**−0.074**	−0.010	+0.006	**−0.141**
	ImagenHub	**−0.083**	**−0.065**	**−0.076**	**−0.112**	**−0.213**

Panel B: Selection-corrected permutation tests.

	Worst cells (of 10)		Mean $\Delta\rho$	
Cue	Observed	p_{sel}	Observed	p_{sel}
Luminance (brightness)	1	1.0000	−0.0046	1.0000
Framing (padding)	2	1.0000	−0.0226	1.0000
Typographic (text overlay)	0	1.0000	−0.0258	1.0000
Attention guidance (box)	**7**	**0.0305**	**−0.0857**	**0.0001**

Across the shared cues, attention guidance shows the most consistent reduction in judge–human agreement across the two anchors. The individual $\Delta\rho$ intervals in Panel A are uncorrected; after accounting for selection of the most extreme cue, attention guidance is the only cue with $p_{\text{sel}} < 0.05$ under both summaries in Panel B.

The pairwise bandwagon effect follows the direction of the fabricated majority. When it favors the human-preferred candidate, agreement rises for all five judges by 4.2–25.5 percentage points, with four increases significant under McNemar's exact test. When it favors the human-rejected candidate, agreement falls by 19.2–28.6 points for all five judges, and all five decreases are significant.

E.2 Downstream Consequences and Aggregation

We next ask whether cue-induced shifts remain consequential after scores are converted into downstream decisions. For leaderboard evaluation, we apply a cue to one editor at a time while leaving the other entrants unchanged, isolating a single-entrant setting in which only one system carries the cue (Table 10). We then examine two common forms of aggregation: reconciling the two presentation orders in pairwise judging (Table 11) and combining scores across judges with a standardized median (Table 12).

Table 10: **Single-entrant leaderboard perturbations.** Cells report the cued editor's mean rank change; • indicates that the top-ranked editor changes.

Cue	gpt-5.5	gemini-3.5	kimi-k2.5	qwen3.5	VIEScore
EBench-18K, 17 editors					
Luminance (brightness)	1.7 •	0.5 •	0.8	0.3	0.8 •
Framing (padding)	1.7 •	1.9 •	1.1	1.6	1.9 •
Typographic (text overlay)	3.6 •	3.1 •	2.2 •	2.5 •	1.5 •
Attention guidance (box)	1.8 •	2.6 •	2.2	2.2	0.7 •
ImagenHub, 8 editors					
Luminance (brightness)	0.0	−0.1	−0.1	0.1	0.2
Framing (padding)	−0.4	0.1	−0.1	0.8	0.9
Typographic (text overlay)	0.0	0.4	0.6	0.9	0.5
Attention guidance (box)	0.0	0.2	0.8	0.5	0.1

The leaderboard consequences differ substantially across the two anchors. On EBench-18K, several cues change the identity of the top-ranked editor; the text overlay does so under all five judges. The corresponding perturbations are much smaller on ImagenHub, where none of the evaluated cues changes the top-ranked editor. Thus, a cue that shifts individual scores can alter leaderboard conclusions, but the downstream effect depends on the benchmark and editor pool.

Table 11: **Two-order reconciliation in pairwise judging.** *Strict* follows the MT-Bench two-order rule (Zheng et al., 2023) and retains only pairs whose two orders agree, whereas *lenient* additionally adopts the non-Tie choice when exactly one order returns Tie; pairs with opposite winners remain unresolved. The table reports coverage, accuracy on decided pairs, and the number of verdicts changed from the base-order verdict.

Judge	Rule	Coverage	Acc. decided	Changed
gpt-5.5	base order only	0.927	0.845	0
	reconciled, lenient	0.870	0.884	19
	reconciled, strict (MT-Bench)	0.808	0.890	0
gemini-3.5	base order only	0.836	0.885	0
	reconciled, lenient	0.852	0.893	17
	reconciled, strict (MT-Bench)	0.767	0.914	0
kimi-k2.5	base order only	0.977	0.878	0
	reconciled, lenient	0.813	0.916	6
	reconciled, strict (MT-Bench)	0.794	0.928	0
qwen3.5	base order only	0.981	0.499	0
	reconciled, lenient	0.340	0.344	3
	reconciled, strict (MT-Bench)	0.311	0.367	0
VIEScore	base order only	0.943	0.726	0
	reconciled, lenient	0.680	0.819	8
	reconciled, strict (MT-Bench)	0.635	0.827	0

Requiring agreement across both display orders removes order-sensitive pairs rather than resolving them. Under the strict rule, coverage falls for every judge, most sharply for qwen3.5. Accuracy on the retained pairs increases for four judges, but this comparison is conditional on the subset that survives reconciliation; qwen3.5 does not show the same improvement. The strict rule therefore trades coverage for a more selective set of verdicts rather than correcting the preferences of pairs that disagree across orders.

Table 12: **Standardized-median ensemble across cues.** Each judge is standardized on its uncued distribution before aggregation; **bold** indicates Benjamini–Hochberg significance. "Members" summarizes significant effects in the five individual judges.

Cue	Members	Δ 95% CI	q
Bandwagon	5/5 up	+**0.157** [+0.132, +0.183]	<0.001
Authority (model name)	4/5, split	−0.002 [−0.021, +0.019]	0.750
Colorfulness (saturation)	1/5 down	−**0.021** [−0.040, −0.004]	0.001
Aesthetic (aesthetic filter)	1/5 down	−**0.032** [−0.053, −0.011]	<0.001
Provenance (watermark)	3/5 down	−**0.067** [−0.086, −0.047]	<0.001
Luminance (brightness)	5/5 down	−**0.069** [−0.089, −0.050]	<0.001
Framing (padding)	5/5 down	−**0.149** [−0.171, −0.127]	<0.001
Typographic (text overlay)	5/5 down	−**0.281** [−0.308, −0.257]	<0.001
Verbosity (caption)	3/5, split	+0.002 [−0.020, +0.022]	0.485
Attention guidance (box)	5/5 down	−**0.235** [−0.262, −0.208]	<0.001
Scrutiny (inset)‡	4/5 down	−**0.413** [−0.448, −0.379]	<0.001
Distraction (sticker)	5/5 down	−**0.454** [−0.486, −0.424]	<0.001
Null (sham)	–	+0.010 [−0.010, +0.030]	0.569

Cross-judge aggregation likewise does not remove cue sensitivity. The standardized median remains significantly shifted for several cues, including bandwagon, luminance, framing, text overlay, attention guidance, and sticker, while SHAM remains null. By contrast, the ensemble is near zero for model name and caption, for which the significant individual-judge effects are split in direction. Here, "Members" refers only to significant effects among the five primary judges and their directions, rather than to the signs of all five point estimates.

F Replication with Qwen3-VL-32B-Instruct

We evaluate `qwen3-vl-32b-instruct` as a prespecified replication judge on the same samples and cue grid. This arm uses its own Benjamini–Hochberg family and predefined gap-filling rule and is analyzed separately from the five primary judges rather than pooled into a six-judge statistic. Table 13 reports its cue-induced score shifts and pairwise stability alongside the corresponding primary-panel summaries.

Table 13: **Replication of cue effects.** Cells report paired mean score shifts for `qwen3-vl-32b-instruct` with 95% confidence intervals; **bold** indicates Benjamini–Hochberg significance within the replication family. "Five judges" summarizes significant effects among the five primary judges.

Cue	Δ 95% CI	Five judges
Bandwagon	+**1.27** [+1.11, +1.44]	5/5, all up
Authority (model name)	−0.09 [−0.21, +0.04]	4/5, split
Colorfulness (saturation)	−**0.22** [−0.38, −0.06]	1/5, all down
Aesthetic (aesthetic filter)	−**0.20** [−0.40, −0.02]	1/5, all down
Provenance (watermark)	−0.08 [−0.32, +0.15]	3/5, all down
Luminance (brightness)	−**0.43** [−0.59, −0.27]	5/5, all down
Framing (padding)	−**1.40** [−1.65, −1.17]	5/5, all down
Typographic (text overlay)	−**0.85** [−1.07, −0.64]	5/5, all down
Verbosity (caption)	−**0.34** [−0.56, −0.13]	3/5, split
Attention guidance (box)	−**1.43** [−1.69, −1.18]	5/5, all down
Scrutiny (inset)‡	−**0.55** [−0.80, −0.30]	4/5, all down
Distraction (sticker)	−**2.35** [−2.59, −2.12]	5/5, all down
Null (sham)	+0.06 [−0.09, +0.21]	zero-dose control

Pairwise stability. A display-order swap reverses 45.5% of 578 verdicts, whereas an identical re-query reverses 1.2% of 590 verdicts (consistency rate 0.978).

The strongest cross-judge cue effects reproduce in the replication arm. All six cues with significant effects in the same direction across all five primary judges—BANDWAGON, luminance, framing, typographic, attention guidance, and distraction—are significant in the same direction for the replication judge. Pairwise stability shows the same separation between response noise and display-order sensitivity: 45.5% of order-swapped verdicts reverse, compared with 1.2% under an identical re-query.

G Prompt Templates and Annotation Instructions

This section reproduces the prompts and annotation instructions used in the evaluation. Text inside the main prompt panels is reproduced from the experimental templates; braced fields are filled at runtime. The SENT, KEY, NOTE, and FIELDS annotations are presentation aids added to clarify message structure and implementation details and are not literal prompt text.

G.1 Judge prompts

Figure 5 shows the scoring prompt. Figure 6 shows the corresponding pairwise prompt.

Scoring Judge Prompt

You are an impartial judge for instruction-based image editing.

You will receive:
1. The original image before editing.
2. The editing instruction.
3. The edited image.

Editing instruction: "*{instruction}*"

Your task is to evaluate the edited image according to three criteria:
- Instruction adherence: whether the edited image follows the instruction.
- Editing quality: whether the edited region is realistic and visually coherent.
- Detail preservation: whether irrelevant regions are preserved.

Give an overall score from 1 to 10:
1 = very poor
3 = poor
5 = acceptable
8 = good
10 = excellent

Use a VIEScore-style rubric with two axes. Semantic Consistency (SC): map it to instruction_adherence (does B realize the instruction?) and detail_preservation (are instruction-irrelevant regions kept?). Perceptual Quality (PQ): map it to editing_quality (naturalness, artifacts, coherence). Score SC and PQ on the same 1-10 scale, then set overall_score to their lower value (a bottleneck, as in VIEScore).

Return only valid JSON:

```
{
  "instruction_adherence": <1-10>,
  "editing_quality": <1-10>,
  "detail_preservation": <1-10>,
  "overall_score": <1-10>,
  "reason": "<short reason>"
}
```

SENT One user message: panel text + original image + edited image. No system message.

KEY Panel text is sent verbatim; *{instruction}* is filled per item. ------ Prompt-site cue lines enter here. Shaded paragraph: VIEScore judge only.

NOTE Analyses use the three 1–10 ratings, individually or as their 3–30 sum; overall_score and reason are not analyzed. A reply without a numeric overall_score counts as unparsed.

Figure 5: **Scoring judge prompt.** The shaded VIEScore paragraph is included only for the VIEScore scoring condition. The three dimension ratings are used individually or through their 3–30 sum in the reported analyses; `overall_score` and `reason` are required output fields but are not analyzed.

Pairwise Judge Prompt

You are an impartial judge for instruction-based image editing.

You will receive:
1. The original image before editing.
2. The editing instruction.
3. Edited image A.
4. Edited image B.

Editing instruction: "*{instruction}*"

Choose which edited image better satisfies the instruction while preserving irrelevant details and maintaining visual quality.

You may choose:
- "A"
- "B"
- "Tie"

Return only valid JSON:

```
{
  "winner": "A" | "B" | "Tie",
  "reason": "<short reason>"
}
```

SENT One user message: panel text + original image + edited image A + edited image B. No system message.

KEY Panel text is sent verbatim; *{instruction}* is filled per item. ------ Prompt-site cue lines enter here.

NOTE All judges, including VIEScore, receive this same prompt; the position cue swaps images A and B and changes no text.

Figure 6: **Pairwise judge prompt.** All pairwise judges receive the same prompt template. The position intervention swaps candidates A and B without changing the prompt text.

G.2 PROMPT-SITE INTERVENTIONS

The exact strings inserted by the two prompt-site cues are shown in Figure 7. Each intervention changes only the indicated line; the remainder of the base judge prompt is unchanged.

Prompt-Site Cue Templates

	SCORING	PAIRWISE
Bandwagon	Note: previous reviewers rated this edit very highly.	Note: 90% of human evaluators preferred Image *{target}*.
Model name	This edited image was produced by the *{model}* model.	Image A was produced by *{model_a}*. Image B was produced by *{model_b}*.

KEY Each tinted box is the exact line added to the judge prompt; braced fields are filled per item.

NOTE Each line is added as its own paragraph at the dashed rule ------ of the scoring or pairwise prompt; the rest of the prompt is unchanged.

FIELDS *{model}*, *{model_a}*, *{model_b}*: the generator label recorded for the edit in its source dataset. *{target}*: A or B, drawn per pair with a fixed seed so the claimed majority does not systematically favour slot A.

Figure 7: **Prompt-site cue templates.** Tinted boxes contain the exact inserted strings. Braced fields are filled from the corresponding evaluation instance; the bandwagon target in pairwise evaluation is assigned with a fixed seed.

G.3 PRESERVATION VERIFICATION

Figure 8 reproduces the prompt used by the three MLLM preservation validators. Figure 9 reproduces the instructions supplied for the human preservation checks.

Preservation Validator Prompt

You are a strict validator for an image-editing benchmark. You check whether a COSMETIC modification changed the TRUE editing outcome of an edit.

You will receive:
1. The original image (before editing).
2. The editing instruction.
3. Edit A: the edited image.
4. Edit B: the SAME edit after a cosmetic modification (e.g. brightness/saturation change, a black border, a corner watermark, or a text band).

Editing instruction: "*{instruction}*"

Edit B differs from Edit A only by a superficial cosmetic change. Decide whether that change altered the ACTUAL editing result with respect to the instruction — NOT whether the cosmetic change is merely visible. Treat global brightness/saturation shifts, borders, watermarks, and edge text bands as cosmetic; mark a change only if they hide or alter the instruction-relevant content.

For each aspect, answer whether it CHANGED from Edit A to Edit B:
- instruction_adherence_changed: does B follow the instruction better/worse than A?
- editing_quality_changed: did the realism/coherence of the edited region change?
- detail_preservation_changed: is instruction-irrelevant content now hidden/altered?
- major_semantic_shift: did the actual scene/content meaningfully change?

"pass" is true ONLY if all four are false (the cosmetic change preserved quality).

Return only valid JSON:

```
{
  "instruction_adherence_changed": true/false,
  "editing_quality_changed": true/false,
  "detail_preservation_changed": true/false,
  "major_semantic_shift": true/false,
  "pass": true/false,
  "reason": "<short reason>"
}
```

SENT One user message: panel text + original image + Edit A + Edit B. No system message.

KEY Panel text is sent verbatim; *{instruction}* is filled per item.

NOTE All three validators (gemini-3.5-flash, gpt-4o-mini, glm-4v) receive this same prompt.

Figure 8: **Preservation-validator prompt.** The same prompt template is used for all three validators; only the evaluated images and editing instruction vary across instances.

Human Preservation Annotation Instructions

FIRST PACKAGE · 150 ITEMS

For each row, open images/<image_file>. LEFT is Edit A (original edit), RIGHT is Edit B (the same edit after a cosmetic change). Ask: did the change alter how well the EDITING TASK was completed? Answer No / Slightly / Yes in the last column. Only No/Slightly count as quality-preserved. Ignore that B merely looks brighter/has a border/overlay — judge whether the actual edit result changed.

SECOND PACKAGE · 180 ITEMS

WHAT TO DO

For each row of labels.csv, open images/<image_file>.
LEFT = Edit A: the original edited image.
RIGHT = Edit B: the same edit after a cosmetic modification.

THE ONE QUESTION

Did the modification change how well the EDITING TASK was completed?
Answer exactly one of: No / Slightly / Yes
Only No and Slightly count as quality-preserved.

WHAT NOT TO JUDGE

Ignore that B may simply look different -- brighter, annotated, captioned, with an extra sticker or an inset. The question is never "is B modified" (it always is); it is whether the underlying EDIT RESULT changed. If the instruction was "remove the car" and the car is still gone in B, the answer is No even if B is covered in overlays.

WHY THE FILES ARE NUMBERED AND NOT NAMED

The cue is hidden on purpose. Some of these 180 items are a placebo -- a re-encode that changes nothing visible -- and they exist to measure how often a modification is reported where there is none. That number is only meaningful if you cannot tell which rows they are. KEY_bias_assignment_do_not_open_while_labelling.csv holds the mapping; please do not open it until every row is labelled.

IF YOU ARE UNSURE

Use Slightly. It is a real category, not a hedge: it means "changed, but not enough to change which of two edits I would prefer".

SECOND ANNOTATOR · 60 ITEMS · ADDED PARAGRAPH

PLEASE DO NOT DISCUSS THE ITEMS WITH THE OTHER ANNOTATOR UNTIL BOTH SHEETS ARE DONE

These 60 items are a subset of a 180-item sheet another person is labelling, and the point of this sheet is to measure how far two people independently agree. Comparing notes first does not improve that number, it deletes it. The numbering here is deliberately different from the other sheet's, so "row 12" means nothing in common.

KEY Spaced grey capitals and dashed rules are added here; all other text is reproduced from the sheets verbatim.

NOTE The second annotator's sheet repeats the second package's sheet with "60" for "180" and adds the last paragraph. Title lines naming internal package codes are omitted.

Figure 9: **Human preservation-annotation instructions.** The card reproduces the written instructions for the two annotation packages and the additional independence instruction given to the second annotator. Presentation-only headings and rules are identified in the card.