Do Image–Text Metrics Respect Semantic Invariances?

Amit Agarwal Hitesh Laxmichand Patel Meizhu Liu Jyotika Singh
Karan Dua Hansa Meghwani Matthew Rowe Michael Avendi
Yassi Abbasi Tao Sheng Sujith Ravi Dan Roth
Oracle AI
Correspondence: amit.h.agarwal@oracle.com

Abstract

Reference-free image-to-text evaluators are now standard for scoring image-caption alignment, yet it is unclear whether they respect semantic invariances. We present an invariance probe on five popular evaluators (CLIPScore, PAC-S, UMIC, FLEUR, and a deterministic LLM judge) under semantics-preserving perturbations along three axes- spatial (flips, context-preserving repositioning, light rotations), object (scale, category), and socio-linguistic framing (cultural/economic adjectives with neutral and length-matched controls). Across curated slices of three detection datasets and three caption evaluation suites, we find consistent non-semantic sensitivities, where benign spatial edits and simple phrasing changes shift scores by ≈6-9% on average, and for systems separated by just 0.7%, these shifts can cause ranking flips in upto ∼37% of cases, particularly under spatial changes. A small human study also supports this finding and confirms that annotators generally judge perturbed pairs as equally correct, so these shifts reflect metric behavior rather than semantic change. We further propose invariance-calibrated scoring, a post-hoc adjustment that roughly halves median absolute sensitivity while retaining correlation with learned caption evaluators.

1 Introduction

Reference-free captioning evaluators are now standard in multimodal research in which they scale without human references and often show strong correlations with human judgments (e.g., CLIPSCORE (Hessel et al., 2021), PAC-S (Sarto et al., 2023a)). Yet such aggregate correlations can mask how evaluators behave under semantics-preserving changes, which are alterations to the layout, object factors, or wording that do not affect the core meaning of the content. A metric may appear stable on average while responding systematically to vertical flips, repositioning that preserves the meaning of the scene, or socially marked adjectives that do not alter the depicted content. When systems are separated by sub-percent gaps, these sensitivities can invert rankings and confound fairness conclusions. In deployed NLP stacks, metric outputs drive downstream decisions ranging from model selection and reward shaping to retrieval ranking and the scoring of natural-language renditions of structured outputs (Singh, 2023; Meghwani et al., 2025; Singh et al., 2025), so evaluator reliability is a practical concern beyond any single benchmark.

We take a *metric-centric invariance audit*: rather than correlating to references or annotators, we maintain a fixed meaning and apply controlled probes across three families (*spatial*, *object*, and *socio-linguistic*) to test whether scores remain stable. We instantiate this on three image sources (COCO (Lin et al., 2014), OpenImages (Kuznetsova et al., 2020), Objects365 (Shao et al., 2019)) and five evaluators spanning embedding similarity and learned caption assessment (CLIPSCORE, PAC-S, UMIC (Lee et al., 2021), FLEUR (Lee et al., 2024), and one deterministic judge). A human validation study confirms that, for almost all paired items, annotators judge both versions acceptable and equally good, so systematic score changes reflect evaluator behavior rather than semantic drift.

We introduce *invariance-calibrated scoring*, a post-hoc adjustment that subtracts per-prompt nuisance sensitivity estimated from invariance families. Under a tight correlation-preservation constraint with learned caption evaluators, it roughly halves median absolute sensitivity on average (largest on spatial probes) while retaining evaluator utility and reducing ranking-flip risk. The contribution is diagnostic, not a claim that one metric objective is correct. Whether evaluators are used for alignment or typicality, the magnitude and asymmetry of the shifts we document warrant disclosure and the opt-in mitigation we propose. Our

Findings of the Association for Computational Linguistics: ACL 2026, pages 39089–39116
July 2-7, 2026

contributions can be summarized as:

- A unit-test framework that audits reference-free captioning evaluators for *semantic invariance* across spatial, object, and socio-linguistic families.
- A cross-source, cross-metric study revealing consistent, practically meaningful sensitivities ($\approx$ 6-9% under benign spatial changes) and linking them to leaderboard instability via an intuitive flip-risk functional.
- A practical mitigation, *invariance-calibrated scoring*, that reduces non-semantic sensitivity without retraining.

2 Related Work

Reference-free metrics have reshaped evaluation for vision-language models, whose dataset and application landscape has expanded rapidly (Pattnayak et al., 2024), offering scalable alternatives to reference-based measures such as CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016). Among them, CLIPScore has been widely adopted, as it computes the alignment between image and caption using pre-trained CLIP embeddings and often correlates better with human judgments than traditional metrics. This has led to its use in selection and reward pipelines (Cho et al., 2023) and analyses of grounding behavior (Barraco et al., 2022). To mitigate documented weaknesses, PAC-S and PAC-S++ (Sarto et al., 2023a, 2025) improve robustness to redundancy and noise. Complementary families such as HICE-S (Zeng et al., 2024) and ensembling approaches like ECO (Jeong et al., 2024) and BRIDGE (Sarto et al., 2023b) combine signals or refine representations to stabilize alignment. Surveys map the space of vision–language evaluation (Zhang et al., 2024), while critiques examine when automatic scores capture meaning rather than surface regularities (Ross et al., 2024). Most work reports aggregate correlations, leaving unclear how metrics behave under semantics-preserving perturbations. Evidence suggests these failures arise from limitations of CLIP-like embeddings and transformer reasoning (Yuan et al., 2026), motivating interpretable alternatives such as DCSMs (Kang et al., 2025). Complementary lines of work benchmark model-side robustness of vision–language systems under targeted perturbations and context shifts, and propose fine-grained scores for multimodal reasoning quality (Agarwal et al., 2025a; Patel et al., 2025; Agarwal et al., 2025b); our audit instead targets the evaluator itself.

A growing body of work also interrogates fairness, accessibility, and framing. Studies probe whether reference-free metrics reflect linguistic intent or user needs (Ahmadi and Agrawal, 2024; Kasai et al., 2022), highlight accessibility gaps (Zur et al., 2024), and propose fairness-oriented designs such as FLEUR (Lee et al., 2024). Parallel efforts also curate multicultural vision–language resources to counter Eurocentric defaults in benchmarks and evaluation, for instance SEA-VL for Southeast Asia (Cahyawijaya et al., 2025). Distinct from prior work, we present a *metric-centric* unit-test audit that keeps core meaning fixed while varying image structure (e.g., spatial position, object scale) and language (e.g., framing, word choice), yielding actionable diagnostics for bias-aware design and reporting. A parallel line of work studies LVLMs as judges for vision–language tasks and examines their reliability and biases (Zhang et al., 2023; Chen et al.; Hwang et al., 2025); our deterministic judge is representative of this family, and our framework treats any such judge as an additional scoring function to be audited rather than as a gold standard.

3 Methodology

We propose an *invariance audit* for captioning-oriented, reference-free vision-language evaluators where the principle is simple, if an image-caption pair's meaning is unchanged, a reliable evaluator should be stable. Rather than relying on aggregate correlations, we apply controlled, semantics-preserving perturbations along spatial, object, and socio-linguistic axes while holding semantics fixed.

We adopt an *evaluation-centric* notion of semantic invariance, in which a perturbation is semantics-preserving when human annotators judge the caption-image relation as equally acceptable for caption-quality scoring. Under this definition, an evaluator used for alignment or caption quality should not materially shift. The validation study bears this out, with both versions majority-acceptable in 97.3% of paired items and majority-tied preferences in 96.6% (Appendix A.11). The scope is deliberate, we do not claim that sensitivity to typicality is inherently a defect of a representation. But when these metrics are deployed for model selection, leaderboards, or reward pipelines, undisclosed sensitivity to nuisance factors produces

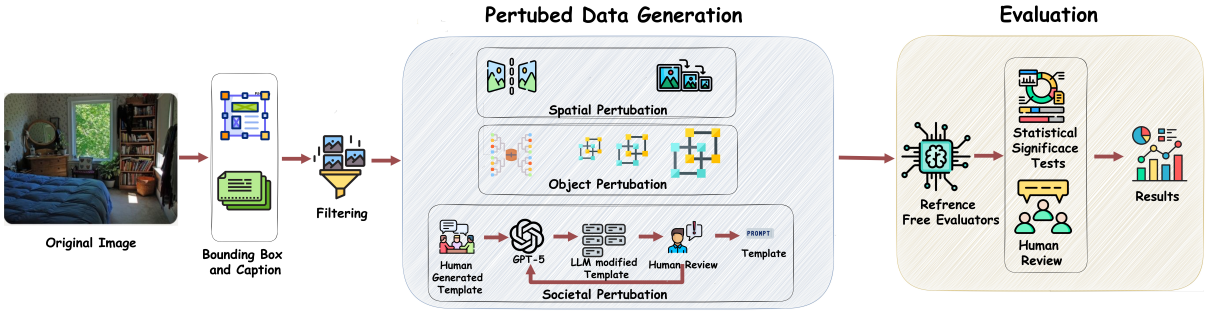

Figure 1: Pipeline Overview. Starting from curated single-object image–caption pairs, we construct matched spatial, object, and socio-linguistic variants and quantify reference-free evaluators sensitivity with paired comparisons.

measurement volatility and ranking instability, regardless of whether one ultimately reads it as a bug or a feature of the encoder.

Metrics under audit. We study five captioning-centric, reference-free evaluators spanning embedding similarity (CLIPSCORE, PAC-S) and learned caption quality (UMIC, FLEUR), plus one deterministic LLM-judge variant with a fixed rubric to triangulate with judge-style practice.[1] The framework is evaluator-agnostic where adding a new scoring function $S(x, c)$, for instance an open-source LVLM-as-judge, only requires implementing a scoring interface.

3.1 Datasets and Curation

We audit three image sources: **MS-COCO 2017 val**, **OpenImages V7 val**, and **Objects365 val**, and we replicate key probes on three caption-evaluation suites (**Flickr8k-CF**, **Pascal-50S**, and **COMPOSITE**) for external validity. For the caption suites we follow the official image pools to avoid domain shift. Across all sources, we use a shared curation protocol and harmonize labels to a single taxonomy.

Single-object slice. From each large-scale source we select images with one prominent instance and high-quality boxes: (i) a single dominant object (with a maximum Intersection over Union (IoU) overlap of < 0.1 with other objects, where IoU measures the ratio of intersection area to the union area between two bounding boxes); (ii) clear visibility; (iii) compatibility with planned transforms (no truncation after repositioning/rotation); (iv) box area sufficient to stratify size. We bin normalized coverage (box area / object area) into seven thresholds: 0-10, 10-20, 20-35, 35-50, 50-70, 70-90, 90-100%. Per-source balance tables (category $\times$ size bin) appear in Appendix A.2. This controlled single-object, templated-caption regime is a deliberate choice to maximize internal validity such that each perturbation targets a specific spatial, object, or socio-linguistic factor while the depicted content remains fixed. External validity under this design is addressed in Section 5.4 (caption-evaluation suites, naturally occurring captions, and a multi-object probe), across which the main conclusions persist.

Category harmonization. We map dataset labels into a shared taxonomy- *person*, *animal*, *vehicle*, *furniture*, *kitchen*, *sports*, *electronics*, *indoor*, *outdoor*. Ambiguous classes are resolved by majority usage; dataset composition and mapping details are reported in Appendix A.2.

Caption templates and lexical families. Captions are programmatically generated from human-designed templates to preserve semantics while varying framing. Controlled synthesis of this kind, where content-bearing factors are parameterized while surrounding structure is held fixed, is increasingly used for training and evaluation data in multimodal pipelines (Dua et al., 2025; Agarwal et al., 2024). We use simple declarative forms such as *"There is a [object]."* and *"There is a [adjective] [object]."*; minimal spatial descriptors are added only when needed for disambiguation so that wording does not encode the downstream transforms. Socio-linguistic modifiers cover *cultural*, *economic*, *gender*, and *emotion* families, as well as a small set of occupational and socio-political descriptors (e.g., *local/foreign*, *immigrant/citizen*) used in targeted probes. Each family includes *neutral* baselines

[1] Reference-based metrics such as CIDEr and SPICE measure agreement with references and are not the focus of this invariance audit.

(e.g., *typical*, *plain*) and *length-matched* variants to control for phrasing length and syntax. The length-matching process ensures that each variant has a similar number of characters or tokens, preventing any bias due to differences in length. For a subset of analyses, we also form simple intersectional combinations (e.g., cultural modifiers applied to person vs. non-person objects) to probe whether sensitivities depend on the underlying category. Further details on screening and lexicons are detailed in the Appendix A.3).

3.2 Perturbation Axes

We construct semantics-preserving perturbations to isolate non-semantic sensitivities while keeping the described content fixed. Each probe is evaluated as a paired contrast against the unperturbed version. Perturbations are grouped into three families -

Spatial. We apply vertical and horizontal flips, *context-preserving repositioning*, and light in-plane rotations ($\pm 10°$). For repositioning, the segmented dominant object is translated within the *original background* at constant scale to four anchors (TL, TR, BL, BR); translations that would collide with boundaries or occlude salient regions are resampled. We also include a *Gaussian-blur control* ($\sigma \in \{1.0, 2.0\}$) on the original image to decouple size from texture/detail loss. To rule out compositing artifacts, we quantify background-change and boundary-seam indicators for repositioned images and verify that the findings persist after filtering artifact-heavy cases (Appendix A.8).

Object. We analyze scores across the seven coverage bins and the harmonized taxonomy while holding captions fixed, probing how object scale and category influence metric behavior independent of wording.

Societal. We replace neutral captions with culturally, economically, and socially marked variants (e.g., *American*/*European*, *cheap*/*expensive*) plus gender and emotion adjectives, chosen to alter socio-linguistic framing while leaving scene semantics unchanged (e.g., *African bed* vs. *American bed* for the same pictured bed). All comparisons include *neutral* and *length-matched* controls so that differences reflect framing rather than phrasing length. We analyze aggregate effects and stratified slices (person vs. non-person; Appendix B.3), and complement templated probes with naturally occurring captions (Appendix A.12); perturbation-axis details are in Appendix A.4.

Human validation of semantics preservation. We validate that our perturbations are semantics-preserving for humans via a small study on $N_{\text{human}} = 480$ paired items from the curated slice (3 annotators/item; Appendix A.11), each comparing two versions of the same example (image perturbation with a shared caption, or caption perturbation with a shared image). A majority judges *both* versions acceptable in 467/480 items (97.3%), and the majority preference is a tie in 464/480 (96.6%). Excluding the rare one-version-acceptable cases (9/480; 1.9%) changes median $\%\Delta$ by at most 0.3 points and never flips effect directions. We therefore treat the perturbations as semantics-preserving where systematic score shifts reflect evaluator sensitivity rather than perceived mismatch.

3.3 Evaluation Protocol

All images are scored with CLIPScore, PAC-S, UMIC, FLEUR, and one deterministic judge under a shared preprocessing pipeline. We use paired contrasts wherever applicable (flip vs. original; neutral vs. modified; TL vs. BR) and report: (i) medians with 95% BCa bootstrap CIs (10k resamples, seed=2025); (ii) normality checks via Shapiro-Wilk; paired t-test when both samples are approximately normal, else Wilcoxon signed-rank; (iii) Kruskal-Wallis for multi-level factors (bins, categories, modifier families) with Holm-adjusted pairwise tests; (iv) effect sizes via Cliff's δ for paired deltas. For comparability to caption evaluators, we also track Spearman/Kendall correlation against UMIC/FLEUR.

Reporting conventions. For perturbation analyses, we report the *Median Relative Change*,

$$\%\Delta \;=\; 100 \times \frac{S_{\text{pert}} - S_{\text{orig}}}{S_{\text{orig}}},$$

aggregated by the median across image-caption pairs, with per-evaluator 95% BCa bootstrap CIs. Positive $\%\Delta$ means the evaluator scored *higher* under the perturbation; larger $|\%\Delta|$ indicates *worse invariance* (it is not a quality gain).

3.4 Risk of Ranking Flip (RRF)

Small but systematic sensitivities can reorder near-tied systems. For a fixed evaluator S and a perturbation family $\mathcal{T}$ (e.g., vertical flips), suppose two captioning systems A and B each produce a caption for every image x. Let $S_A(x, t)$ and $S_B(x, t)$ denote the scores that S assigns to the outputs of A

and B on image x after applying transform $t \in \mathcal{T}$ (with t=id the unperturbed case), and define the pointwise score gap

$$\Delta_S(x,t) \;=\; S_A(x,t) - S_B(x,t).$$

We define the *risk of ranking flip* as

$$\mathrm{RRF}_S(A,B;\mathcal{T}) = \Pr_{x\sim\mathcal{D},\, t\sim\mathcal{T}}\big[\operatorname{sign}\Delta_S(x,t) \neq \operatorname{sign}\Delta_S(x,\mathrm{id})\big], \quad (1)$$

i.e., the probability that the ordering between A and B under S changes after applying a semantics-preserving perturbation. We estimate RRF using paired bootstraps over images and transforms, and report it as a function of a fixed average score gap d (e.g., d=0.7% on COCO) between A and B on the unperturbed data. We operationalize near-tied instability via a fixed-gap stress test at d=0.7%; full definition and estimation are in Appendix B.5. Intuitively, RRF answers when given two near-tied systems, how often would their leaderboard order flip under benign changes such as flips or cultural wording?

3.5 Invariance-Calibrated Scoring

To reduce nuisance sensitivity *without retraining* an evaluator, we post-hoc adjust scores using sensitivities measured under our invariance probes. Let $S(x,c)$ be the raw score that an evaluator assigns to an image-caption pair (x,c), and let $\mathbb{T}$ denote the set of perturbation families we audit (e.g., spatial, linguistic modifiers).

For a given family $\mathcal{T} \in \mathbb{T}$, we write $(x^{(t)}, c^{(t)})$ for the perturbed pair obtained from (x,c) under transform $t \in \mathcal{T}$ (with t=id the original). We quantify the *sensitivity* of S for (x,c) along $\mathcal{T}$ as

$$\Delta_S(x,c;\mathcal{T}) = \mathrm{median}_{t\in\mathcal{T}}\left|S\left(x^{(t)},c^{(t)}\right) - S(x,c)\right|$$

the median absolute score change across transforms in that family. The calibrated score is then

$$\hat{S}(x,c) \;=\; S(x,c) \;-\; \lambda \sum_{\mathcal{T}\in\mathbb{T}} w_{\mathcal{T}}\, \Delta_S(x,c;\mathcal{T}),$$

with non-negative weights $w_{\mathcal{T}}$ (default uniform) and a global strength parameter $\lambda \geq 0$.

Intuitively, $\hat{S}$ subtracts nuisance sensitivity estimated from our invariance tests. On a held-out dev set we sweep and choose the smallest value that substantially reduces median absolute sensitivity while keeping Spearman correlation with learned caption evaluators (UMIC, FLEUR) within 0.01 of the uncalibrated metric. We report both sensitivity reductions and changes in RRF.

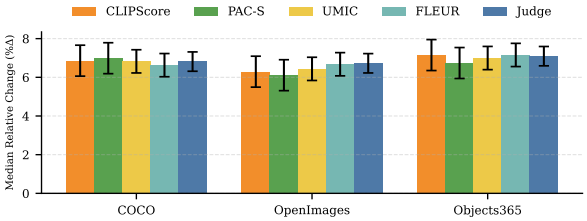

Figure 2: **RQ1a (Vertical flips).** Median %Δ (95% CIs) across evaluators and datasets; positive values indicate higher post-flip scores (orientation sensitivity).

4 Experiments

We structure our experiments to find answers to targeted questions for the study:
RQ1 (Spatial invariance) How do scores change under flips, repositioning, and light rotations?
RQ2 (Object sensitivity). How do scores vary with object scale and category?
RQ3 (Socio-linguistic framing). How do simple cultural and economic adjectives affect scores relative to neutral controls?
RQ4 (Cross-dataset behavior). Do the trends generalize across metrics and caption suites?
RQ5 (Ranking impact). For near-tied systems, how often do perturbations invert their ranking?
RQ6 (Mitigation). Does invariance-calibrated scoring reduce sensitivities and flip risk while preserving correlation with human-aligned caption evaluators?

We apply the perturbation families from Section 3.2 to five evaluators across COCO, OpenImages, and Objects365, replicating key trends on the three caption suites. Each test pairs an unperturbed image-caption with a minimally altered counterpart, so any systematic change reflects metric sensitivity rather than genuine mismatch.

5 Results

We investigate patterns and answer RQ1-RQ6 across multiple datasets and metrics. Additional experiments and full details are in Appendix B.

5.1 RQ1: Spatial Invariance

Vertical flips (RQ1a). Across COCO, OpenImages, and Objects365, all five evaluators increase under vertical flips by ≈6–8% (median %Δ),

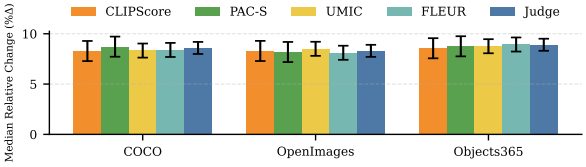

Figure 3: **RQ1b (Repositioning).** Median %Δ (BR−TL) across evaluators and datasets (95% CIs). Repositioning induces sizable shifts (≈7–9%); BR>TL.

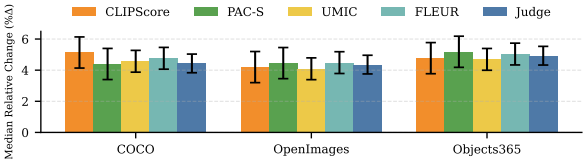

Figure 4: **RQ1b (Rotation).** Median %Δ for ±10° rotations (95% CIs). Smaller than repositioning (≈4–6%), yet consistent across evaluators and datasets.

with 95% CIs strictly above zero for every evaluator–dataset pair (Fig. 2). Mixed-effects meta-estimates confirm this, with small-to-moderate paired effect sizes (Table 3). These shifts are large enough to invert near-tied systems (quantified in §5.5), and the human-validation check (below) rules out perceived semantic change as an explanation.

Context-preserving repositioning and rotations (RQ1b). Absolute position in the same scene also matters: moving the dominant object from top-left (TL) to bottom-right (BR) yields the largest spatial deltas (≈7–9%) across evaluators and datasets (Fig. 3), while light in-plane rotations (±10°) are smaller but reliable (≈4–6%; Fig. 4). Across-quadrant differences are significant (Kruskal–Wallis; Holm-adjusted pairwise tests confirm BR>TL; full per-dataset breakdowns in Table 4).

Two checks rule out the explanation that these shifts simply track perceptual quality loss. First, annotators return a majority tie in relative preference on 96.6% of paired items and judge both versions acceptable for 97.3% of items (Appendix A.11); fewer than 2% show a majority-unacceptable version, so spatial perturbations do not make one version systematically worse to humans. Second, a Gaussian-blur control ($\sigma \in \{1.0, 2.0\}$) applied to the *original* image reduces detail and naturalness but leaves object position and framing unchanged, and produces markedly smaller shifts than context-preserving repositioning (subsection 5.2, Figure 5). If naturalness or detail loss were the primary driver, blur would shift scores at least as much as repositioning; it does not. We therefore attribute the observed TL→BR and ±10° deltas to non-semantic positional sensitivity rather than to human-perceived correctness or naturalness degradation.

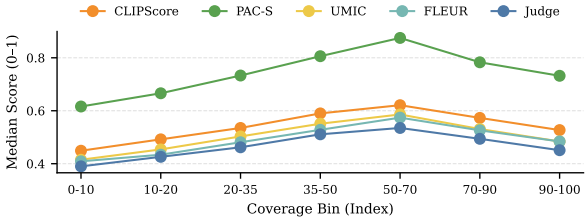

Figure 5: **RQ2a (Size).** Median score by object coverage bin across evaluators. All metrics peak at 50–70% coverage and drop at the smallest/largest bins.

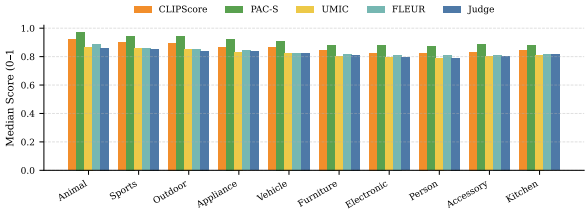

Figure 6: **RQ2b (Category).** Median score by category on COCO across evaluators. Category differences are stable across evaluators; *Animal* > *Vehicle* > *Person*.

5.2 RQ2: Object Sensitivity

Scale (RQ2a). Across evaluators, scores follow an *inverted-U*: performance peaks at 50-70% coverage and declines at very small/very large targets (Fig. 5), with magnitudes larger for embedding similarity than learned caption evaluators. On COCO, extremes are penalized by ≈6–9% relative to the mid-band for embedding metrics and less for learned evaluators; Kruskal–Wallis with Holm-adjusted pairwise tests confirms mid-band > extremes. A blur control on the *original* image yields much smaller shifts, ruling out texture/detail loss as the sole driver.

Category (RQ2b). Under the harmonized taxonomy, macro-median scores differ by several points across categories (Fig. 6), with *Animal* near the top and *Person* consistently lower; the spread persists after adjusting for size, implying that category mix can shift headline metrics (Appendix B.2).

5.3 RQ3: Socio-linguistic Framing

Cultural modifiers (RQ3a). Relative to neutral phrasing, cultural adjectives induce a consistent ordering with small–to–moderate magnitude on COCO (Fig. 7). Embedding-similarity evaluators show the strongest effects: for CLIPScore, *African* yields the largest negative me-

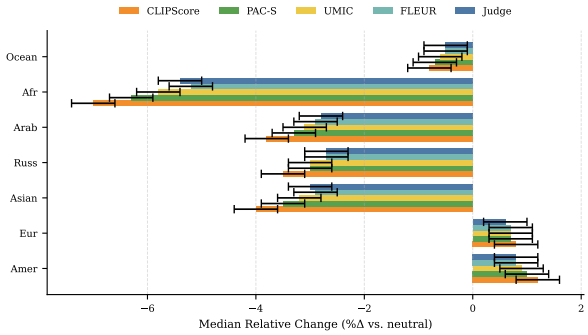

Figure 7: **RQ3a (Cultural).** Median %Δ vs. neutral on COCO across evaluators (95% CIs). Consistent ordering with larger shifts in embedding-similarity metrics.

dian change (≈-7%), while *American/European* are mildly positive (≈+1%); UMIC/FLEUR and the JUDGE track the same order at lower magnitude. Because the image content is unchanged, these shifts reflect sensitivity to non-visual cultural framing rather than scene semantics. A complementary analysis on naturally occurring captions (Appendix A.12) shows qualitatively similar behavior when we neutralize or swap cultural descriptors in human-written captions: for example, CLIPSCORE still assigns lower scores on average to captions containing *African* than to matched captions containing *American/European*.

Economic modifiers (RQ3b). For non-person objects, *cheap* is mildly positive while *expensive* is consistently negative across evaluators (Fig. 8). On COCO, CLIPSCORE medians are ≈+2.0% (*cheap*) vs. ≈-6.2% (*expensive*); PAC-S shows ≈+1.2% vs. ≈-5.8%. UMIC/FLEUR/JUDGE display smaller but directionally aligned shifts. These results indicate that even simple economic framing can systematically alter metric judgments despite identical visuals. When we apply the same procedure to naturally occurring captions that mention price (e.g., "an expensive car" vs. "a car" or "a cheap car"; Appendix A.12), we again observe that the *median paired score shift* against the neutral caption follows *cheap* > *neutral* > *expensive* across evaluators.

Embedding-side analysis. To understand why CLIP-based metrics exhibit these socio-linguistic asymmetries, we analyze the geometry of the CLIP/PAC-S text encoders (Appendix A.13). We construct a simple "valence" direction in CLIP text space from clearly positive vs. negative adjectives and project each socio-linguistic modifier (and short adjective-noun phrases) onto this direction. We compute, across modifiers a, the Spearman correlation between (i) each modifier's projection onto a CLIP text-space "valence" direction, $s(a) = \langle e(a), v_{\text{val}} \rangle$, and (ii) the modifier's induced median score shift $\widetilde{\Delta}(a)$ (median % Δ vs. the neutral baseline when inserting a into an otherwise unchanged caption). We obtain $\rho \approx 0.7$, meaning modifiers that are more "negative" along this direction tend to produce more negative score shifts, consistent with framing effects being partly driven by text-encoder geometry. This suggests that pretraining-induced valence structure in the text encoder contributes to the observed effects.

Interpretation. Our socio-linguistic perturbations use curated adjectives as controlled probes: the image and denoted object are fixed while framing varies along cultural, economic, gender, and affective dimensions. The probes can in principle reveal two distinct failure modes. (a) A *uniform groundedness penalty*, in which the evaluator lowers scores for any adjective it cannot verify against the image; drops would then be symmetric across modifiers of equal groundedness and could be defended as a reasonable evaluation philosophy. (b) *Modifier-identity sensitivity*, in which the evaluator responds asymmetrically to the identity of the modifier, beyond what groundedness alone would predict. What we observe is (b): on the same generic images, where *American bed* and *African bed* are equally ungrounded, CLIPSCORE yields $\approx +1\%$ for *American/European* but $\approx -7\%$ for *African* (subsection 5.3, Figure 7). This asymmetry indicates differential priors under controlled equivalence rather than a principled groundedness penalty. The same ordering persists in naturally occurring, and therefore more plausibly grounded, captions when we neutralize or counterfactually swap the modifier (Appendix A.12), and the text-encoder valence analysis (Appendix A.13) shows these effects are systematically predicted by geometry in the CLIP text space. Annotators do not penalize the modifiers either: both versions are judged acceptable for $>97\%$ of items and preferences are majority-tied for 96.6% (Appendix A.11). We read these results as diagnostics of evaluator behavior, not claims about real-world groups; the practical concern is that modifier-identity sensitivity propagates into model selection, ranking, and reward pipelines, where it destabilizes near-tied rankings (subsection 5.5) regardless of whether one frames the sensitivity as a bug or a feature.

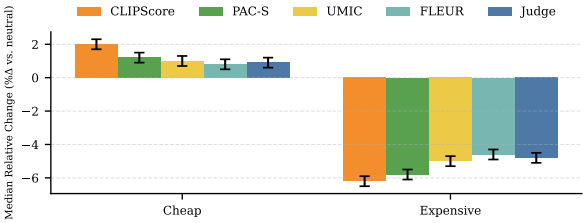

Figure 8: **RQ3b (Economic).** COCO: median %Δ vs. neutral across evaluators (95% CIs). *Cheap* is mildly positive, while *expensive* is consistently negative.

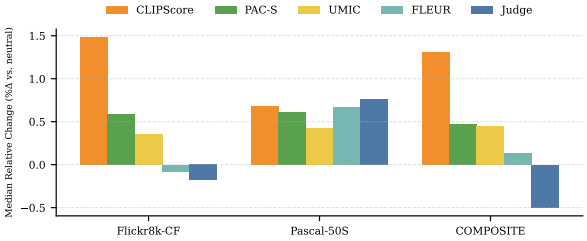

Figure 9: **RQ4 (External-Cultural).** Results on external corpora (avg. *American/European/Asian/African*) highlight persistent effects with reduced magnitude.

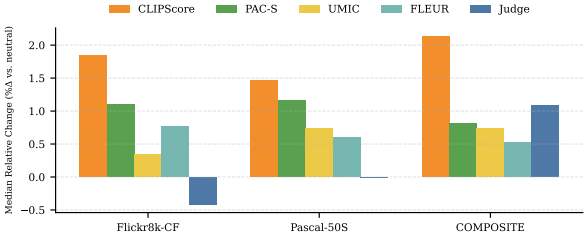

Figure 10: **RQ4 (External-Economic).** Median %Δ vs. neutral on external corpora (avg. *cheap/expensive*). *Cheap* is mildly positive while *expensive* is negative.

5.4 RQ4: Cross-Dataset behavior

Patterns from RQ1–RQ3 persist across evaluators and transfer to three caption-evaluation suites (Flickr8k-CF, Pascal-50S, COMPOSITE). On these external corpora, vertical flips remain *positive* for all evaluator–corpus combinations: medians cluster in the mid–single digits for embedding-similarity metrics and are attenuated yet positive for UMIC/FLEUR/JUDGE. Our judge is instantiated from a GPT-5 model (Appendix A.5), the fact that it exhibits similar qualitative spatial sensitivities as CLIPSCORE/PAC-S, albeit at lower magnitude, suggests that these invariance failures are not confined to shallow similarity measures.

Socio-linguistic framing also replicates: the cultural ordering (*American/European* slightly positive; *African* most negative) and the economic pattern (*cheap* mildly positive; *expensive* negative) hold across corpora (Figs. 9–10), again with smaller magnitudes for learned evaluators and the LLM judge. These results indicate that the sensitivities we document are not an artifact of a single dataset or metric family. Extended results and direction agreement are reported in Appendix B.4.

External validity summary. Three complementary checks indicate that the sensitivities in RQ1-RQ3 are not artifacts of our templated, single-object regime. **(i) Caption-evaluation suites.** RQ1-RQ3 replicate on Flickr8k-CF, Pascal-50S, and COMPOSITE with consistent effect directions across all evaluator–corpus pairs (Figs. 9–10; Appendix B.4, Tables 9–11). **(ii) Naturally occurring captions.** On a probe of 732 human-written captions from MS-COCO and the three suites that already contain our socio-linguistic adjectives (Appendix A.12), neutralization and counterfactual swaps reproduce the same ordering as the templated probes; for example, CLIPSCORE penalizes captions containing *African* relative to matched *American/European* variants by roughly -5%, and *expensive* relative to neutral by -4% to -6%. **(iii) Multi-object pilot.** A 200-image MS-COCO probe with 2-4 prominent objects (Appendix A.10) shows that spatial effect directions match the single-object slice, with slightly attenuated magnitudes, higher variance, and no qualitative reversals. We therefore treat the audit as a controlled unit-test whose conclusions are supported, though not exhausted, by external-validity evidence on natural captions, multi-object scenes, and standard caption-evaluation suites.

5.5 RQ5: Risk of Ranking Flip

We quantify how often a near-tied leaderboard can invert under semantics-preserving changes using the flip-risk functional RRF (Section 3.4), which directly connects to model selection and deployment risk. For a fixed gap d=0.7% on COCO, spatial perturbations (especially *context-preserving repositioning*) produce the largest risks, followed by *cultural* framing; *economic* modifiers and light *rotations* are smaller but clearly non-zero (Fig. 11, Appendix B.5). For CLIPSCORE, median RRF is about 28% (vertical flips), 36% (repositioning), and 18% (rotation); socio-linguistic framing is smaller but material (≈16% cultural, ≈12% economic; Table 13). PAC-S is similar or slightly higher on spatial axes, while learned evaluators (UMIC/FLEUR) and the judge are consistently lower (e.g., ≈23–27% for repositioning) but still show substantial flip risk. When top systems differ by sub-percent to ∼1%, leaderboard order can

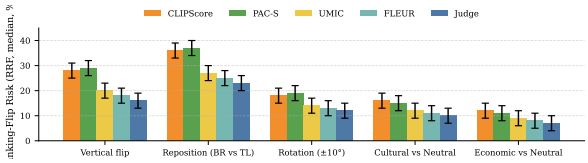

Figure 11: **RQ5 (Flip risk).** Median RRF (%) by perturbation family (95% CIs). Ranking instability is highest under spatial probes, especially repositioning.

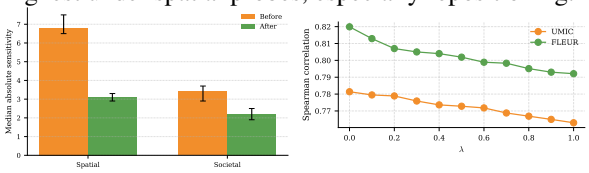

Figure 12: **RQ6 (Calibration).** Sensitivity by axis before calibration (left) and utility retention via dev-split Spearman vs. λ (right); $\lambda^\star$ chosen under ϵ=0.01 w.r.t. UMIC/FLEUR.

invert on roughly one in three evaluation pairs under benign, semantics-preserving changes, most acutely for object repositioning.

5.6 RQ6: Invariance-calibrated scoring

We apply invariance-calibrated scoring to down-weight non-semantic sensitivity while preserving *agreement* with established caption evaluators. With a tight agreement constraint (ϵ=0.01 vs. UMIC/FLEUR on a dev split), calibration reduces *median absolute sensitivity* by roughly half on average across spatial and societal axes; gains are largest on spatial probes (repositioning, flips) and smaller yet consistently non-zero on socio-linguistic probes. Embedding-similarity evaluators (CLIPSCORE, PAC-S) benefit most, while UMIC/FLEUR start lower and show smaller reductions. Figure 12 shows before/after sensitivity.

Calibration also narrows socio-linguistic framing gaps: the median absolute cultural shift across modifiers drops from ≈3.4% to ≈2.2% (35%), and the spatial gap from ≈6.8% to ≈3.1% (54%). Finally, calibration reduces RRF by lowering the dominant driver: spatial sensitivity, yielding double-digit RRF drops for repositioning (Table 16). Computationally, calibration is a one-time offline sweep over a fixed set of perturbation variants per dev item; at inference it adds only constant-time post-processing on top of the base metric (Appendix B.6). To further validate beyond correlations, we evaluate calibrated scores on human-labeled caption-preference suites; calibration preserves (and in some cases slightly improves) pairwise preference accuracy (Appendix B.7, Table 17). Taken together, invariance-calibrated scoring offers a lightweight mitigation: it reduces nuisance sensitivity and ranking volatility while retaining agreement with standard evaluators and human-preference fidelity.

The calibration is extensible by design: new invariance axes can be added to $\mathcal{T}$ without retraining the evaluator, with the concrete deployment recipe in Appendix B.6. It therefore complements, rather than replaces, ongoing work to broaden the set of audited invariances, and it cannot correct axes that have not yet been instantiated.

6 Conclusion

Reference-free captioning evaluators are convenient, but they are not invariant to semantics-preserving changes. Our unit-test framework shows consistent sensitivities across three axes (spatial: orientation, position, light rotation; object: scale, category; socio-linguistic framing: cultural/economic adjectives) and across datasets and metric families, even though a human validation study indicates that these perturbations preserve caption correctness for annotators. These shifts are large enough to destabilize near-tied leaderboards and to induce substantial risk of ranking-flip under benign spatial transforms. We frame these findings as a diagnostic contribution rather than a prescription of a single correct metric objective, whether one treats reference-free evaluators as alignment-focused or typicality-focused, the magnitude and asymmetry of the observed shifts, particularly on socio-linguistic axes, are large enough to warrant explicit disclosure and the kind of post-hoc mitigation we propose.

Practical guidance. To make evaluations more robust and comparable, we recommend (i) reporting *paired* statistics with CIs and effect sizes; (ii) adding category/scale-stratified summaries; (iii) using neutral and length-matched controls for wording; and (iv) adopting simple invariance checks (flip, reposition, rotation) as pre-submission unit tests for metrics and production systems. Our *invariance-calibrated scoring* provides a post-hoc option that reduces non-semantic sensitivity and RRF while preserving utility correlations.

Outlook. Future evaluators should encode spatial awareness, normalize for scale/category composition, and temper language-side framing effects, encouraging bias-aware multimodal evaluation.

7 Limitations

Our audit is positioned as a controlled unit-test rather than an exhaustive evaluation of reference-free captioning in the wild. The curated single-object slice and templated captions are a deliberate internal-validity choice: by fixing syntax and isolating content to one dominant instance, we can attribute paired score changes to the targeted spatial, object, or socio-linguistic factor rather than to multi-object interactions such as attribute binding, occlusion, co-occurrence, or relational ambiguity. We complement this controlled core with three external-validity checks (replication on three caption-evaluation suites, a multi-object sanity probe, and a socio-linguistic probe on naturally occurring human-written captions), across which the qualitative effects persist. These checks are supporting evidence, not a complete substitute for open-ended evaluation: a full audit on densely compositional scenes, relational captions, and free-form prompting is left to future work.

We study five widely used reference-free evaluators (CLIPSCORE, PAC-S, UMIC, FLEUR, and one deterministic GPT-5-based judge). While this spans embedding-, learned-, and judge-style scoring, it does not cover the full evaluator space (e.g., alternative CLIP backbones, other learned metrics, or multiple judge prompts/configurations). Our calibration is evaluated on the metrics and probe families in scope; extending it to additional metrics or tasks is a natural next step. The proposed calibration targets invariance-style nuisance factors (e.g., orientation/position and prompt framing) and is not intended to remove object-scale or category effects, which likely reflect systematic evaluator preferences and dataset/model composition.

Our socio-linguistic probes use curated adjective sets as *instruments* to test invariance under controlled wording changes. They are intentionally limited and should not be interpreted as statements about real-world groups or cultures; we report these effects as properties of the evaluators under audit.

Despite these constraints, the findings are consistent across three large sources and persist under neutral and length-matched controls.

Ethical Considerations

We use publicly available vision-language datasets (MS-COCO, OpenImages, Objects365, Flickr8k-CF, Pascal-50S, COMPOSITE) under their published licenses and official splits. We release only image identifiers, prompt templates/lexicons, and derived evaluator outputs/analysis artifacts (not the underlying images). Since these corpora may contain sensitive or biased content, our goal is to diagnose evaluator behavior under controlled, semantics-preserving perturbations rather than to validate dataset labels or make claims about depicted subjects.

Our socio-linguistic probes vary caption framing using curated adjective lists (cultural, economic, gender, emotion) while holding the image constant; reported gaps characterize evaluator sensitivity and should not be interpreted as statements about real-world groups. Where human validation is used, it involves low-risk judgments of caption acceptability for synthetic perturbations. We analyze failure modes of widely used reference-free metrics, but we do not propose a single "correct" evaluator. There is a risk that our calibration recipe or flip-risk functional could be misused to optimize leaderboard position without improving reliability or fairness.

References

Eslam Abdelrahman, Pengzhan Sun, Li Erran Li, and Mohamed Elhoseiny. 2024. Imagecaptioner2: Image captioner for image captioning bias amplification assessment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 20902–20911.

Amit Agarwal, Srikant Panda, Angeline Charles, Hitesh Laxmichand Patel, Bhargava Kumar, Priyaranjan Pattnayak, Taki Hasan Rafi, Tejaswini Kumar, Hansa Meghwani, Karan Gupta, and Dong-Kyu Chae. 2025a. MVTamperBench: Evaluating robustness of vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18804–18828, Vienna, Austria. Association for Computational Linguistics.

Amit Agarwal, Hitesh Laxmichand Patel, Srikant Panda, Hansa Meghwani, Jyotika Singh, Karan Dua, Paul Li, Tao Sheng, Sujith Ravi, and Dan Roth. 2025b. RCI: A score for evaluating global and local reasoning in multimodal benchmarks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 138–157, Suzhou (China). Association for Computational Linguistics.

Amit Agarwal, Hitesh Laxmichand Patel, Priyaranjan Pattnayak, Srikant Panda, Bhargava Kumar, and Tejaswini Kumar. 2024. Enhancing document AI data generation through graph-based synthetic layouts. *International Journal of Engineering Research & Technology (IJERT)*, 13(10). ArXiv:2412.03590.

Saba Ahmadi and Aishwarya Agrawal. 2024. An examination of the robustness of reference-free image captioning evaluation metrics. *Preprint*, arXiv:2305.14998.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic Propositional Image Caption Evaluation. In *European Conference on Computer Vision (ECCV)*, pages 382–398. Springer.

Michele Barraco, Marcella Cornia, Stefano Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. The unreasonable effectiveness of clip features for image captioning: an experimental analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4662–4670.

Samuel Cahyawijaya, Holy Lovenia, Joel Ruben Antony Moniz, Tack Hwa Wong, Mohammad Rifqi Farhansyah, Thant Thiri Maung, Frederikus Hudi, David Anugraha, Muhammad Ravi Shulthan Habibi, Muhammad Reza Qorib, Amit Agarwal, Joseph Marvin Imperial, Hitesh Laxmichand Patel, and 1 others. 2025. Crowdsource, crawl, or generate? creating SEA-VL, a multicultural vision-language dataset for Southeast Asia. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18685–18717, Vienna, Austria. Association for Computational Linguistics.

Santiago Castro, Amir Ziai, Avneesh Saluja, Zhuoning Yuan, and Rada Mihalcea. 2024. Clove: Encoding compositional language in contrastive vision-language models. *arXiv preprint arXiv:2402.15021*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, 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 *Forty-first International Conference on Machine Learning*.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2023. Fine-grained image captioning with clip reward. *Preprint*, arXiv:2205.13115.

Karan Dua, Hitesh Laxmichand Patel, Puneet Mittal, Ranjeet Gupta, Amit Agarwal, Praneet Pabolu, Srikant Panda, Hansa Meghwani, Graham Horwood, and Fahad Shah. 2025. FlexDoc: Parameterized sampling for diverse multilingual synthetic documents for training document understanding models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1500–1521, Suzhou (China). Association for Computational Linguistics.

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

Yusuke Hirota, Yuta Nakashima, and Noa Garcia. 2022. Quantifying societal bias amplification in image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13450–13459.

Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. Infometic: An informative metric for reference-free image caption evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3171–3185.

Yerin Hwang, Dongryeol Lee, Kyungmin Min, Taegwan Kang, Yongil Kim, and Kyomin Jung. 2025. Fooling the lvlm judges: Visual biases in lvlm-based evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23197–23216.

Kiyoon Jeong, Woojun Lee, Woongchan Nam, Minjeong Ma, and Pilsung Kang. 2024. Technical report of nice challenge at cvpr 2024: Caption reranking evaluation using ensembled clip and consensus scores. *Preprint*, arXiv:2405.01028.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. Tiger: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2141–2152.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. 2025. Is clip ideal? no. can we fix it? yes! *arXiv preprint arXiv:2503.08723*.

Jungo Kasai, Keisuke Sakaguchi, Lavinia Dunagan, Jacob Morrison, Ronan Le Bras, Yejin Choi, and Noah A. Smith. 2022. Transparent human evaluation for image captioning. *Preprint*, arXiv:2111.08940.

Yongil Kim, Yerin Hwang, Hyeongu Yun, Seunghyun Yoon, Trung Bui, and Kyomin Jung. 2023. Pr-mcs: Perturbation robust metric for multilingual image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12237–12258.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: an open dataset of user preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. 2020. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. UMIC: An Unreferenced Metric for Image Captioning via Contrastive Learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 220–226. Association for Computational Linguistics.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. 2020. Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the first workshop on evaluation and comparison of NLP systems*, pages 34–39.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. FLEUR: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3732–3746, Bangkok, Thailand. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Zheng Ma, Changxin Wang, Yawen Ouyang, Fei Zhao, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2024. Cobra effect in reference-free image captioning metrics. *arXiv preprint arXiv:2402.11572*.

Kazuki Matsuda, Yuiga Wada, and Komei Sugiura. 2024. Deneb: A hallucination-robust automatic evaluation metric for image captioning. In *Proceedings of the Asian Conference on Computer Vision*, pages 3570–3586.

Hansa Meghwani, Amit Agarwal, Priyaranjan Pattnayak, Hitesh Laxmichand Patel, and Srikant Panda. 2025. Hard negative mining for domain-specific retrieval in enterprise systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1013–1026, Vienna, Austria. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Dong Huk Park, Samaneh Azadi, Xihui Liu, Trevor Darrell, and Anna Rohrbach. 2021. Benchmark for compositional text-to-image synthesis. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Hitesh Laxmichand Patel, Amit Agarwal, Srikant Panda, Hansa Meghwani, Karan Dua, Paul Li, Tao Sheng, Sujith Ravi, and Dan Roth. 2025. PCRI: Measuring context robustness in multimodal models for enterprise applications. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 195–214, Suzhou (China). Association for Computational Linguistics.

Priyaranjan Pattnayak, Hitesh Laxmichand Patel, Bhargava Kumar, Amit Agarwal, Ishan Banerjee, Srikant Panda, and Tejaswini Kumar. 2024. Survey of large multimodal model datasets, application categories and taxonomy. *Preprint*, arXiv:2412.17759.

Haoyi Qiu, Zi-Yi Dou, Tianlu Wang, Asli Celikyilmaz, and Nanyun Peng. 2023. Gender biases in automatic evaluation metrics for image captioning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8358–8375.

Candace Ross, Melissa Hall, Adriana Romero Soriano, and Adina Williams. 2024. What makes a good metric? evaluating automatic metrics for text-to-image consistency. *Preprint*, arXiv:2412.13989.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023a. Positive-augmented contrastive learning for image and video captioning evaluation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6914–6924.

Sara Sarto, Manuele Barraco, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2023b. Positive-augmented contrastive learning for image and video captioning evaluation. *Preprint*, arXiv:2303.12112.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Bridge: Bridging gaps in image captioning evaluation with stronger visual cues. In *European Conference on Computer Vision*, pages 70–87. Springer.

Sara Sarto, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2025. Positive-augmented contrastive learning for vision-and-language evaluation and training. *Preprint*, arXiv:2410.07336.

Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Jing Li, Xiangyu Zhang, and Jian Sun. 2019. Objects365: A large-scale, high-quality dataset for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8425–8434.

Harman Singh, Pengchuan Zhang, Qifan Wang, Mengjiao Wang, Wenhan Xiong, Jingfei Du, and Yu Chen. 2023. Coarse-to-fine contrastive learning in image-text-graph space for improved vision-language compositionality. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 869–893.

Jyotika Singh. 2023. *Natural Language Processing in the Real World: Text Processing, Analytics, and Classification*, 1st edition. Chapman & Hall/CRC Data Science Series. Chapman and Hall/CRC, Boca Raton, FL.

Jyotika Singh, Weiyi Sun, Amit Agarwal, Viji Krishnamurthy, Yassine Benajiba, Sujith Ravi, and Dan Roth. 2025. Can LLMs narrate tabular data? an evaluation framework for natural language representations of text-to-SQL system outputs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 883–902, Suzhou (China). Association for Computational Linguistics.

Brandon Abreu Smith, Miguel Farinha, Siobhan Mackenzie Hall, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. Balancing the picture: Debiasing vision-language datasets with synthetic contrast sets. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. 2022. Measuring representational harms in image captioning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 324–335.

Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 15903–15935.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Michelle Yuan, Weiyi Sun, Amir H. Rezaeian, Jyotika Singh, Sandip Ghoshal, Yao-Ting Wang, Miguel Ballesteros, and Yassine Benajiba. 2026. Barriers to discrete reasoning with transformers: A survey across depth, exactness, and bandwidth. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1966–1978, Rabat, Morocco. Association for Computational Linguistics.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2023. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*.

Youngsik Yun and Jihie Kim. 2024. Cic: a framework for culturally-aware image captioning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1625–1633.

Zequn Zeng, Jianqiao Sun, Hao Zhang, Tiansheng Wen, Yudi Su, Yan Xie, Zhengjue Wang, and Bo Chen. 2024. Hicescore: A hierarchical metric for image captioning evaluation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, page 866–875. ACM.

Jing Zhang, Jing Huang, Sheng Jin, and Shuicheng Lu. 2024. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. To appear.

Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. 2023. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*.

Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14830–14840.

Amir Zur, Elisa Kreiss, Karel D'Oosterlinck, Christopher Potts, and Atticus Geiger. 2024. Updating clip to prefer descriptions over captions. *Preprint*, arXiv:2406.09458.

A Appendix

A.1 Extended Related Work

Recent advancements in image captioning evaluation highlight the need to move beyond traditional n-gram based metrics eg BLEU(Papineni et al., 2002), CIDEr (Vedantam et al., 2015), SPICE(Anderson et al., 2016), due to their limited semantic fidelity and poor alignment with human judgment. A growing body of research has shifted towards reference-free metrics that leverage deep multimodal representations and integrate explainability and robustness to perturbations. At the same time, bias and fairness in captioning models and their evaluation have gained significant attention, particularly regarding the amplification of societal biases in both caption generation and evaluation processes.

A.1.1 Evaluation Metrics for Image Captioning

Recent advances in image captioning evaluation emphasize moving beyond n-gram metrics such as BLEU(Papineni et al., 2002), CIDEr (Vedantam et al., 2015), and SPICE(Anderson et al., 2016), which often fall short on semantic fidelity and human alignment. Research has increasingly shifted to reference-free evaluators using pretrained multimodal representations, with added focus on interpretability, robustness to perturbations, and fairness, since both captioning models and metrics can encode and amplify societal biases.

Reference-free metrics have now become central to the evaluation of image captioning models. CLIPScore measures semantic image–text alignment using pretrained CLIP embeddings, offering a reference-free evaluation that correlates strongly with human judgment (Hessel et al., 2021). CLoVe extends the vision-language model framework by refining text-image alignment through object and attribute recognition, making it highly relevant for fine-grained evaluations (Castro et al., 2024). PickScore fine-tunes CLIP-H on a large dataset of human-generated images and corresponding preferences, improving alignment with user expectations and reflecting human satisfaction (Kirstain et al., 2023).

Going beyond scalar scoring, InfoMetIC provides more fine-grained analysis by identifying missing or incorrectly described content, offering better interpretability compared to traditional metrics (Hu et al., 2023). ImageReward evaluates the alignment of generated captions with human preferences, directly incorporating human feedback to improve alignment (Xu et al., 2023). Human Preference Score (HPS) fine-tunes the CLIP-L model, focusing on increasing alignment between user-chosen images and textual descriptions, offering a robust metric for subjective quality assessment (Wu et al., 2023).

Robustness to text perturbations has become a key focus. PR-MCS fine-tunes CLIP's text encoder to maintain stability under lexical perturbations, ensuring better performance across variations in input text (Kim et al., 2023). Similarly, DENEB reduces sensitivity to hallucinations by training a transformer-based similarity measure on human-annotated data (Matsuda et al., 2024). Cobra Effect in Reference-Free Metrics warns that optimizing models against imperfect evaluators can lead to inflated scores that do not genuinely improve the semantic alignment between text and images (Ma et al., 2024).

ViLBERTScore evaluates image captioning by using a vision-and-language BERT model to compute textual embeddings for a reference and generated caption. The embeddings are conditioned on the target image, offering a contextual approach to text-image alignment (Lee et al., 2020).Complementarily, BRIDGE introduces a learnable multimodal evaluator that better captures visual evidence and text-image alignment by learning from both visual and textual features(Sarto et al., 2024).

BLIP-ITC and BLIP2-ITC use contrastive learning for image-text alignment, employing cosine similarity to generate more accurate text-to-image retrieval scores (Li et al., 2022, 2023). TIGEr improves evaluation by integrating text-image grounding, better aligning with human judgment than traditional word-based metrics like BLEU, ROUGE, and METEOR (Jiang et al., 2019). CLIP-R-Precision extends R-Precision by incorporating CLIP embeddings for more human-aligned ranking (Park et al., 2021; Xu et al., 2018). NegCLIP refines text-image alignment by improving CLIP's ability to reject irrelevant captions (Yuksekgonul et al., 2023). MosaiCLIP uses scene graphs and graph decomposition to enhance text-image representation, improving alignment by modeling complex relationships between objects (Singh et al., 2023).

A.1.2 Bias, Fairness, and Representational Harm in Captioning

Bias and fairness are increasingly central to captioning and its evaluation. Understanding and Evaluating Racial Biases in Image Captioning documented demographic disparities in caption quality and content (Zhao et al., 2021). Metrics can also be biased: (Qiu et al., 2023) showed CLIPScore may reward gendered language even when inappropriate, reinforcing stereotypes through evaluation. Societal bias amplification work further demonstrates that captioning models can exacerbate biases from training data, motivating careful debiasing strategies (Hirota et al., 2022). ImageCaptioner links these harms to dataset bias (e.g., COCO, Visual Genome) and proposes multimodal protocols to assess bias relative to image content (Abdelrahman et al., 2024). Broader fairness framing emphasizes representational harms such as misrepresentation and omission (Wang et al., 2022), while culturally-aware captioning argues for evaluations that respect diverse contexts and avoid Eurocentric assumptions (Yun and Kim, 2024). Mitigation directions include balanced synthetic contrast sets to reduce spurious correlations in evaluation (Smith et al.).

Together, these threads motivate evaluation that is both robust to perturbations and sensitive to bias. Our work contributes by introducing a unit-test framework that audits captioning metrics under spatial, object, and linguistic perturbations while probing susceptibility to bias and representational harms in realistic settings.

A.2 Datasets, Curation, and Balance

Sources. We use MS-COCO 2017 (val), Open Images V7 (val), Objects365 (val), plus three caption-evaluation suites (Flickr8k-CF, Pascal-50S, COMPOSITE) for external validity (Section 3.1).

Single-object protocol. Images are included when (i) exactly one dominant instance (*max* IoU overlap < 0.1), (ii) minimal occlusion, (iii) transformations do not truncate the instance, and (iv) normalized area defines one of seven size bins (0–10, 10–20, 20–35, 35–50, 50–70, 70–90, 90–100%). Rejections and reasons are logged and released.

Taxonomy mapping. COCO supercategories, Open Images, and Objects365 labels are mapped into *person*, *animal*, *vehicle*, *furniture*, *kitchen*, *sports*, *electronics*, *indoor*, *outdoor*. Ambiguities are resolved by majority use; ties are flagged and released. Distribution across the taxonomy is depicted in Figure 13.

A.3 Caption Templates and Lexicons

Templates. We generate captions from a small set of human-designed templates. Base captions use a simple declarative form: *Base*: "There is a [object].". Attribute captions insert a single adjective: *Attribute*: "There is a [adjective] [object].". Spatial templates add minimal descriptors only when needed for disambiguation (e.g., "There is a [object] on the left.") so that spatial wording does not itself encode the downstream perturbations.

Societal adjective families. Socio-linguistic adjectives are organized into four primary families plus a small socio-political extension: (i) *cultural* (e.g., *American*, *European*, *Asian*, *Arab*, *African*, *Russian*, *Oceanian*); (ii) *economic* (e.g., *cheap*, *expensive*, *luxury*, *budget*); (iii) *gender* (e.g., *male*, *female*, *boy*, *girl*); (iv) *emotion* (e.g., *happy*, *sad*, *angry*); and (v) a small set of *occupational / socio-political* descriptors (e.g., *local*, *foreign*, *immigrant*, *citizen*, *refugee*, *tourist*) used only in targeted probes. Each family includes neutral and length-matched controls such as *typical*, *plain*, or *ordinary*, chosen to match the number of characters and tokens as closely as possible.

We construct adjective–noun pairs by combining these modifiers with compatible object types. For a subset of analyses we build simple intersectional combinations (e.g., *African man*, *American car*, *expensive sofa*) to test whether sensitivities differ between person vs. non-person categories and between object types.

Compatibility screen. We filter ill-formed adjective–noun pairs (e.g., POS conflicts, animacy mismatches, or implausible combinations such as *angry chair*) using a set of heuristics and manual

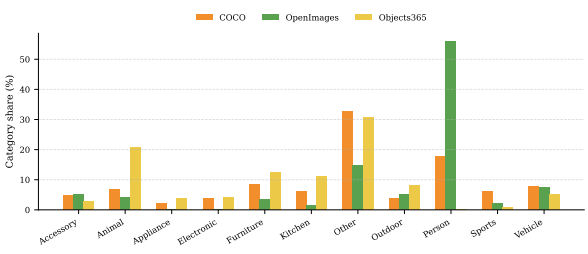

Figure 13: **Category composition of curated slices.** Share (%) of images per harmonized category in COCO/OpenImages/Objects365.

Axis	Perturbations	Probes
Spatial	Flips; context-preserving repositioning; $\pm 10^{\circ}$ rotations; distance-to-center; blur control	Invariance to orientation/layout; texture/detail sensitivity
Object	Seven coverage bins; harmonized categories	Scale effects; category composition
Societal	Cultural/economic/gender/emotion with neutral and length-matched controls	Sensitivity to non-visual framing

Table 1: Perturbations and targeted properties.

spot checks. The compact main subset used in the body is released alongside the *rejected* pairs and their rejection reasons.

A.4 Perturbation Implementations

Table 1 summarizes the perturbation families and the invariance properties they probe and the implementation details are as follows -

Flips and rotations. Vertical/horizontal flips are image-space transforms; rotations use $\pm 5^{\circ}, \pm 10^{\circ}$ with reflection padding and center-preserving resampling.

Context-preserving repositioning. We preserve background and scale, translating the segmented dominant object; collisions/occlusions trigger resampling until a valid placement is found (Algorithm 1).

Repositioning artifact diagnostics. We compute (i) a background-change score as the mean absolute pixel difference outside the union of the source and target object regions, and (ii) a boundary-seam score using edge energy in a thin ring around the pasted mask boundary. We re-run RQ1b after removing the top 5% (and 10%) highest-seam samples; effect directions and medians remain stable.

Distance-to-center proxy. We use coverage bins as a coarse proxy (main text, Fig. 5); a continuous sweep with cubic splines is provided in the released notebooks.

A.5 Extended Experimental Setup

Shared preprocessing. All evaluators use the same resized-crop resolution and normalization; captions pass through the same tokenizer. We log image IDs, seeds, and exact scripts.

CLIPScore / PAC-S. CLIPSCORE computes the cosine of pooled image/text embeddings; PAC-S uses CLIP feature similarity with perceptual calibration. Backbones and pooling choices are recorded.

UMIC / FLEUR. We use publicly released caption-quality evaluators; we record model version/commit and interface limits.

Judge. Our judge is instantiated from a GPT-5 model exposed through an evaluation API. We use a frozen rubric and system prompt that instructs the model to score how well a caption describes an image on a 0–10 scale, with emphasis on factual correctness and relevance. Inference is deterministic (temperature $T{=}0$, fixed seed, no sampling from multiple candidates), and runs are cached by (imageID, caption, rubric) so that repeated uses are bitwise identical. We treat this judge as representative of mainstream LLM-based evaluation practice and report it alongside CLIPSCORE, PAC-S, UMIC, and FLEUR throughout.

Implementation details. All evaluators run deterministically on identical pre-processing; the judge uses a fixed rubric, temperature $= 0$, and a seeded context. Bootstrap uses $10{,}000$ resamples (seed $= 2025$).

A.6 Statistical Procedures and Meta-Analysis

Paired contrasts. Medians of paired differences with 95% BCa bootstrap CIs (10k resamples, fixed seed). Normality via Shapiro–Wilk; two-tailed paired t-test when applicable, otherwise Wilcoxon.

Multi-level factors and multiplicity. Kruskal–Wallis across bins/categories; Holm-adjusted pairwise tests.

Effect sizes and mixed-effects. Cliff's δ for paired deltas. Random-intercept mixed-effects models report β_1 (95% CI) alongside nonparametrics.

A.7 Calibration Ablations and Alternatives

Selecting $\lambda^{\star}$ under a correlation constraint. We choose $\lambda^{\star}$ on a development split by grid search

Algorithm 1 Anchor-Relocation Composition

Require: image I, mask M (dominant object), anchors $A = \{\text{TL}, \text{TR}, \text{BL}, \text{BR}\}$

$P \leftarrow I \odot M$ ▷ object patch

$B \leftarrow I \odot (1 - M)$ ▷ background

for $a \in A$ **do**

  $t \leftarrow \textsc{TranslateCentroidToAnchor}(P, a)$

  while t collides with image boundary **or** occludes salient regions **do**

    $t \leftarrow \textsc{ResampleJitterTowardCenter}()$ ▷ up to K tries

  $I_a \leftarrow B \oplus \textsc{Translate}(P, t)$ ▷ feathered alpha to avoid hard edges

return $\{I_a\}_a$

$\lambda \in \{0, 0.05, \ldots, 1.0\}$, minimizing total non-semantic sensitivity $\sum_{\mathcal{T} \in \mathbb{T}} \text{median}\left|\Delta_{\hat{S}}(\cdot; \mathcal{T})\right|$ subject to correlation preservation: $\text{corr}(\hat{S}, H) \geq \text{corr}(S, H) - \epsilon$ for $H \in \{\text{UMIC}, \text{FLEUR}\}$. The main text (Fig. 12) reports (i) aggregate before/after median absolute sensitivity by axis and (ii) correlation vs. λ with the selected $\lambda^\star$ markers.

Per-metric outcomes. We report for each evaluator: (a) axis-wise median absolute sensitivity (spatial/object/societal) before→after, (b) the selected $\lambda^\star$, and (c) the corresponding Δcorrelation with UMIC/FLEUR.

Ablations. We verify that the calibration behavior is not sensitive to minor design choices:

- **Axis weights.** Uniform weights vs. sensitivity-proportional weights ($w_{\mathcal{T}} \propto$ baseline $\text{median}\,|\Delta|$) yield similar reductions at comparable Δcorrelation.(see `tables/calib_axisweights.csv`).
- **Constraint strength.** We sweep $\epsilon \in \{0.005, 0.01, 0.02\}$; chosen $\lambda^\star$ shifts as expected but sensitivity reductions remain qualitatively stable.
- **Alternative formulations.** We test a ridge-style shrinkage and a quantile-subtraction variant; both reduce sensitivity but are slightly less stable than the default across evaluators.

A.8 Repositioning transform diagnostics

Motivation. Context-preserving repositioning composes a segmented object patch at a new location on the *same* background. Because imperfect masks or blending could introduce low-level artifacts, we run diagnostics to verify that (i) the background is unchanged away from the moved object and (ii) our RQ1b effects persist after removing artifact-heavy composites.

Background preservation (BGΔ). Let I be the original image and I' the repositioned image (RGB normalized to $[0, 1]$). Let M_s be the source object mask in I and M_t the target mask location in I'. We exclude a small neighborhood around either object region and compute mean absolute background change:

$$U = \text{dilate}(M_s \cup M_t, r_{\text{bg}}),$$

$$\text{BG}\Delta(I, I') = \frac{1}{|U^c|} \sum_{p \in U^c} \left\|I(p) - I'(p)\right\|_1,$$

where $\|\cdot\|_1$ sums absolute differences across channels. Low BGΔ indicates background preservation up to negligible interpolation.

Boundary seam strength (SeamE). To quantify potential halo/seam artifacts at the pasted boundary, we measure Sobel gradient energy in a thin ring around the target mask:

$$R = \text{dilate}(M_t, r_2) \setminus \text{dilate}(M_t, r_1),$$

$$\text{SeamE}(I') = \frac{1}{|R|} \sum_{p \in R} G(I'(p)),$$

where $G(\cdot)$ is gradient magnitude (computed per-channel and averaged). We report *relative* seam strength as a ratio $\text{SeamE}(I')/\text{SeamE}(I)$, where $\text{SeamE}(I)$ is computed on an analogous ring around M_s in the original image; ratios near 1 indicate no systematic seam amplification beyond natural object boundaries.

Robustness to artifact filtering and blending. We recompute the RQ1b repositioning effect (median $\%\Delta$ for BR$-$TL with the same BCa protocol as main experiments) after filtering the top $q\%$ highest-scoring samples by BGΔ, by SeamE, or by either. We also re-render a stratified subset using a *feathered alpha* boundary (a 3px linear ramp) to

soften hard cut edges. Across filters and blending choices, effect directions are unchanged and medians remain within overlapping confidence intervals, indicating that the repositioning sensitivity is not driven by a small set of artifact-heavy composites.

Fixed parameters. We use $r_{\text{bg}} = 8$ px, $(r_1, r_2) = (2, 5)$ px for the seam ring, and $q \in \{5, 10\}$ for filtering, fixed *a priori* and applied uniformly across datasets/evaluators. "Feathered alpha" uses a 3px linear ramp at the mask boundary to reduce hard-edge transitions.

A.9 External Validity on Caption Evaluation Suites

We replicate RQ1-RQ3 on Flickr8k-CF, Pascal-50S, and COMPOSITE using their image pools and pairings. To limit domain shift, we apply spatial flips and a reduced cultural/economic subset. Effect directions replicate; magnitudes vary with caption style.

A.10 Multi-object sanity check

Setup. To partially address the gap between our single-object regime and real-world multi-object scenes, we run a small sanity check on a COCO subset with two to four prominent objects. We sample 200 images that pass basic quality filters (no extreme occlusions, reasonably sized instances) and construct captions using the same templated scheme (e.g., "There is a [object].") targeting the dominant object. We then apply the same spatial perturbations as in RQ1 (vertical flips, context-preserving repositioning, light rotations) to the full image while keeping the caption fixed.

Results. Across evaluators, the direction of effects matches the single-object slice: vertical flips and repositioning systematically increase scores, with magnitudes slightly attenuated but within the confidence bands of the single-object results; rotations show smaller but consistent positive shifts. Because the multi-object images introduce additional clutter and interacting objects, variance is higher and some per-category effects become noisier, but we do not observe any reversals in the qualitative patterns.

Interpretation. This probe is deliberately small and not a full multi-object audit, but it suggests that the spatial sensitivities documented in the main text are not an artifact of the single-object regime. A more thorough treatment of cluttered scenes and relational captions is left for future work.

A.11 Human validation of perturbations

Sampling. We annotate $N_{\text{human}} = 480$ paired items from the curated COCO single-object slice, stratified by perturbation family: (1) **Spatial** ($n = 160$): 80 vertical flips, 40 TL↔BR repositionings, and 40 $\pm 10^{\circ}$ rotations. (2) **Object/texture control** ($n = 160$): Gaussian-blur controls applied to images drawn from size extremes (80 smallest-bin, 80 largest-bin); (3) **Socio-linguistic** ($n = 160$): 40 each for cultural, economic, gender, and emotion modifiers. Each item contains two versions of the same underlying example. For spatial/object items we pair the original and perturbed *image* with a shared caption; for socio-linguistic items we use a shared image with a neutral vs. modified caption. The two versions are shown in randomized left/right order.

Annotation protocol. Annotators (English-speaking; 3 per item) answer: (i) **Acceptability (per version).** "How well does this caption describe the image?" with options *incorrect*, *partially correct*, *fully correct*. We map *partially/fully correct* to *acceptable* and *incorrect* to *unacceptable*. (ii) **Relative preference.** "Which is better, or are they equally good?" with options *A*, *B*, *Tie*.

Aggregate outcomes. Using majority vote ($\geq 2/3$), 467/480 items (97.3%) are *acceptable for both* versions; 9/480 (1.9%) are acceptable for only one version; the remainder have no majority or both unacceptable. For relative preference, 464/480 items (96.6%) are majority *tie*; non-tie preferences are roughly balanced between the two versions. Inter-annotator agreement is substantial (Fleiss' $\kappa = 0.63$ for acceptability; $\kappa = 0.59$ for preference).

Effect on reported sensitivities. We perform two complementary robustness checks on the aggregate outcomes. First, we recompute the main spatial and socio-linguistic analyses after excluding the 9 items where only one version is majority-acceptable; across evaluator–dataset pairs, median $\%\Delta$ changes by at most 0.3 percentage points and all effect directions are preserved. Second, as a stricter check, we also remove the 3.1% $(15/480)$ of items for which either version was marked only *partially correct* by the majority of annotators, so

Condition	Ret.	BGΔ ($\times 10^3$)	SeamE ratio	CLIPScore	PAC-S	UMIC	FLEUR	Judge
None (all)	100%	0.25	1.01	8.4	8.7	7.6	7.2	6.6
Drop top 5% SeamE	95%	0.24	0.99	8.3	8.6	7.5	7.1	6.5
Drop top 5% BGΔ	95%	0.18	1.01	8.4	8.7	7.6	7.2	6.6
Drop top 5% either	92%	0.18	0.99	8.2	8.5	7.4	7.0	6.4
Drop top 10% either	85%	0.15	0.98	8.1	8.4	7.3	6.9	6.3
Feathered-alpha subset	20%	0.26	0.96	8.3	8.6	7.5	7.1	6.5

Table 2: **Diagnostics for context-preserving repositioning (RQ1b).** BGΔ is mean absolute background change outside a dilated union of source/target masks (lower is better; values shown $\times 10^3$ for readability). SeamE ratio compares boundary edge energy in I' to the analogous ring in I (ratio ≈ 1 indicates no seam amplification). Right columns report the resulting median %Δ (BR$-$TL) per evaluator after filtering/blending; values remain stable, supporting that RQ1b is not driven by compositing artifacts.

that the remaining set contains items judged *fully correct* on both versions. Recomputing the main analyses on this stricter subset leaves medians, effect directions, and the outcomes of the significance tests (paired Wilcoxon / Kruskal–Wallis with Holm adjustment) unchanged up to the same ≤ 0.3 percentage point tolerance. Together, these two checks indicate that the reported metric sensitivities are not driven by a small set of non-equivalent pairs or by items near the acceptability boundary, and support treating our perturbations as semantics-preserving for humans.

A.12 Socio-linguistic framing probe

Data selection. To test whether the socio-linguistic sensitivities observed with templated captions also appear for naturally occurring language, we search MS-COCO and the three caption-evaluation suites (Flickr8k-CF, Pascal-50S, COMPOSITE) for captions that contain adjectives from our socio-linguistic lexicons (Appendix A.3). We focus on cultural (e.g., *American*, *African*), economic (e.g., *cheap*, *expensive*), gender, and emotion modifiers. After filtering for clear, single-object descriptions and removing duplicates, we obtain 732 base captions across corpora (412 cultural, 196 economic, 124 gender/emotion).

Neutralization and counterfactual rewrites. For each base caption c_{orig} containing a socio-linguistic adjective a modifying an object o, we construct:

- a **neutralized** caption c_{neu} by replacing a with a neutral, length-matched control (e.g., *typical*, *plain*);
- an **alternate** caption c_{alt} where we swap a for an antonym or contrasting modifier from the same family (e.g., *African*$\rightarrow$*American/European*; *cheap*$\rightarrow$*expensive*).

We keep the remainder of the caption identical and reuse the original image. We discard rewrites that become ungrammatical or pragmatically implausible under manual spot checks.

Results. We run the same five evaluators on (x, c_{orig}), (x, c_{neu}), and (x, c_{alt}). For cultural adjectives, CLIPSCORE and PAC-S show the same ordering as in the templated setting: captions containing *African* are scored lower on average than matched captions containing *American/European*, with median relative changes of roughly -5% vs. neutralized variants; *American/European* are mildly positive ($\approx$+0.5–1.0%) relative to neutral. UMIC and FLEUR attenuate magnitudes but preserve the ordering. For economic descriptors, *cheap* captions are again mildly positive (median $\approx$+1–2%) and *expensive* captions negative (median $\approx$-4–6%) relative to their neutralized counterparts, with learned evaluators closer to zero but directionally aligned.

These magnitudes are smaller than those observed under templated captions (Section 5.3, Appendix B.3), reflecting the greater linguistic and contextual variability of real captions, but the qualitative patterns persist. This complementary probe suggests that the socio-linguistic framing sensitivities we document are not an artifact of our templating scheme and can arise in naturally written descriptions as well.

A.13 Text-encoder embedding analysis

To probe whether the socio-linguistic sensitivities documented in Section 5.3 are reflected in the underlying text encoders, we analyze CLIP/PAC-S embeddings for the adjectives used in our cultural, economic, gender, and emotion families.

Setup. We use the same CLIP text backbone used by CLIPSCORE and PAC-S. For each adjective a (e.g., *African*, *American*, *cheap*, *expensive*), we

compute its normalized text embedding $e(a)$ as well as phrase embeddings $e(a, o)$ for simple adjective–noun phrases (e.g., "African bed", "expensive car") where o ranges over a representative subset of objects from our taxonomy. We also construct a simple "valence" direction in text space,

$$\begin{aligned} v_{\text{val}} = \tfrac{1}{n}\big(&e(\text{“good”}) + e(\text{“nice”}) \\ &+ \ldots + e(\text{“beautiful”})\big) \\ &- \tfrac{1}{n}\big(e(\text{“bad”}) \\ &+ \ldots + e(\text{“ugly”}) + e(\text{“terrible”})\big), \end{aligned}$$

and normalize v_{val} to unit length. For each adjective (or phrase), we compute the scalar projection $s(a) = \langle e(a), v_{\text{val}} \rangle$ (or $s(a, o) = \langle e(a, o), v_{\text{val}} \rangle$).

Cultural and economic adjectives. We then correlate these projections with the median relative change $\%\Delta$ induced by the corresponding modifiers when inserted into captions (Tables 7-8). Across the seven cultural adjectives on COCO (American, European, Asian, Arab, African, Russian, Oceanian), we observe a strong monotone relationship between $s(a)$ and the median $\%\Delta$ for CLIPSCORE: Spearman $\rho \approx 0.70$ ($p < 10^{-2}$) at the word level and $\rho \approx 0.78$ for phrase embeddings $e(a, o)$ averaged over objects o. PAC-S exhibits a similar pattern with $\rho \approx 0.66$ (word) and $\rho \approx 0.73$ (phrase). Adjectives with more negative projections on v_{val} (e.g., *African*, *cheap*, negative emotions) systematically yield lower metric scores when used in captions, whereas adjectives with more positive projections (*American*, *European*, *happy*) are associated with higher scores.

For economic modifiers, *cheap* and *expensive* cluster on opposite sides of the valence direction: *cheap* lies slightly above the neutral baseline (*typical*/*plain*), while *expensive* lies below. This matches the metric behavior in Table 8 and Figure 8: *cheap* is mildly positive relative to neutral, whereas *expensive* is strongly negative.

Gender and emotion. Applying the same analysis to gender (*male*/*female*) and emotion (*happy*/*sad*/*angry*) adjectives yields analogous results. For emotion, the projection scores order as *happy* $>$ neutral $>$ *sad*/*angry*, and this ordering matches the median $\%\Delta$ in Figure 15. Correlations are again higher for phrase-level embeddings (e.g., "sad person") than for adjectives alone.

Interpretation. These findings suggest that CLIP's text encoder encodes socio-linguistic descriptors along a latent valence axis in a way that is predictive of how CLIPSCORE/PAC-S respond to them in caption scoring. In other words, the socio-linguistic framing effects documented in Section 5.3 are not arbitrary artifacts of our templating scheme but emerge from systematic structure in the underlying text representation. This supports the view that pretraining-induced priors over cultural, economic, and affective descriptors can directly influence downstream evaluation metrics built on these encoders.

B Extended Results

This appendix provides extended results for all research questions (RQ1–RQ5), including additional probes and per-dataset breakdowns. We document the exact statistical tests (with corrections) and include supporting evidence to substantiate the claims in the main text.

B.1 RQ1: Spatial Invariance

Statistical Procedure. All contrasts are paired at the image level. We report medians of paired differences with 95% BCa bootstrap CIs (10,000 resamples, fixed seed). Normality is screened with Shapiro–Wilk; when non-normality is detected we use Wilcoxon signed-rank (two-sided), otherwise paired t. Across multi-level factors (quadrants), we use Kruskal–Wallis with Holm-adjusted pairwise comparisons. Mixed-effects models use random intercepts for *image* and *category* (REML) and report the perturbation coefficient β_1 with 95% CIs; these complement the nonparametric summaries.

Vertical flips (RQ1a). Per-dataset medians, 95% CIs, Cliff's δ, and mixed-effects meta-estimates appear in Table 3. Across all evaluator–dataset cells, medians are positive with CIs excluding zero, indicating a consistent increase after a vertical flip. Effect sizes are small-to-moderate but practically meaningful given RQ5 (RRF).

Context-preserving repositioning and rotations (RQ1b). Table 4 reports per-dataset medians and 95% CIs for (i) context-preserving relocation from top-left (TL) to bottom-right (BR) at constant scale, and (ii) light in-plane rotations ($\pm 10^{\circ}$). Across quadrants, Kruskal–Wallis detects differences for every evaluator; Holm-adjusted pairwise tests confirm BR $>$ TL. Rotation effects are smaller than repositioning but consistently above zero.

Additional diagnostics. (1) *Horizontal flips* mirror the vertical-flip trend with slightly reduced

Evaluator	COCO	OpenImages	Objects365	Meta β_1 [CI]
CLIPScore	6.9% [6.1,7.7] (δ 0.28)	6.3% [5.5,7.1] (δ 0.28)	7.2% [6.4,8.0] (δ 0.28)	0.069 t[0.061,0.077]
PAC-S	7.0% [6.2,7.8] (δ 0.28)	6.1% [5.3,6.9] (δ 0.28)	6.7% [5.9,7.5] (δ 0.28)	0.070 [0.062,0.078]
UMIC	6.8% [6.2,7.4] (δ 0.20)	6.4% [5.8,7.0] (δ 0.20)	7.0% [6.4,7.6] (δ 0.20)	0.068 [0.062,0.074]
FLEUR	6.6% [6.0,7.2] (δ 0.20)	6.7% [6.1,7.3] (δ 0.20)	7.2% [6.6,7.8] (δ 0.20)	0.066 [0.060,0.072]
Judge	6.8% [6.3,7.3] (δ 0.16)	6.7% [6.2,7.2] (δ 0.16)	7.1% [6.6,7.6] (δ 0.16)	0.068 [0.063,0.073]

Table 3: RQ1a: Vertical flip sensitivity (%Δ). Median [95% CI]; Cliff's δ; mixed-effects β_1 (REML).

Evaluator	COCO	OpenImages	Objects365	Type
CLIPScore	8.3% [7.3,9.3]	8.3% [7.3,9.3]	8.6% [7.6,9.6]	Reposition
CLIPScore	5.1% [4.1,6.1]	4.2% [3.2,5.2]	4.8% [3.8,5.8]	Rotation
PAC-S	8.7% [7.7,9.7]	8.2% [7.2,9.2]	8.8% [7.8,9.8]	Reposition
PAC-S	4.4% [3.4,5.4]	4.5% [3.5,5.5]	5.2% [4.2,6.2]	Rotation
UMIC	8.3% [7.6,9.0]	8.5% [7.8,9.2]	8.8% [8.1,9.5]	Reposition
UMIC	4.6% [3.9,5.3]	4.1% [3.4,4.8]	4.7% [4.0,5.4]	Rotation
FLEUR	8.4% [7.7,9.1]	8.1% [7.4,8.8]	8.9% [8.2,9.6]	Reposition
FLEUR	4.8% [4.1,5.5]	4.5% [3.8,5.2]	5.0% [4.3,5.7]	Rotation
Judge	8.6% [8.0,9.2]	8.3% [7.7,8.9]	8.9% [8.3,9.5]	Reposition
Judge	4.4% [3.8,5.0]	4.4% [3.8,5.0]	4.9% [4.3,5.5]	Rotation

Table 4: RQ1b: Context-preserving repositioning and light rotations (%Δ). Median [95% CI].

magnitude (omitted for space). (2) A *distance-to-center* sweep (score vs. normalized radius with cubic splines) shows monotone improvements toward the bottom-right region, consistent with the quadrant ordering. (3) A *blur control* applied to the original image yields much smaller shifts than repositioning, indicating that the position effect is not explained by texture/detail loss alone.

Mixed-effects interpretation. Random-intercept models (image, category) stabilize estimates across heterogeneous content. Perturbation coefficients β_1 for flips and repositioning are positive with CIs excluding zero across evaluators, matching the paired-bootstrap results. We recommend reporting both the paired median Δ and the mixed-effects β_1 when summarizing spatial sensitivity.

Practical significance. The magnitudes in Table 3 and Table 4 translate into nontrivial *RRF* for near-tied systems (Section 5.5 and Appendix B.5); we therefore advocate pairing headline scores with RRF at a standard gap (e.g., d=0.7%) whenever spatial perturbations are plausible.

B.2 RQ2: Object Sensitivity

Statistical procedure. All summaries are paired at the image level. We report medians with 95% BCa bootstrap CIs (10,000 resamples, fixed seed). For multi-level factors (size bins, categories) we use Kruskal–Wallis with Holm-adjusted pairwise tests. Mixed-effects models (REML) include random intercepts for *image* and *category* and report the effect of the factor of interest; for size we also fit a spline on continuous coverage as a sensitivity check.

Scale (RQ2a). Table 5 gives medians by bin for each evaluator; OpenImages/Objects365 follow a similar inverted-U profile. Relative to the mid-band (50–70%), penalties at the smallest and largest bins are typically ≈6–9% for embedding-based metrics and moderately smaller for learned evaluators. A Gaussian-blur control applied to the *original* image produces much smaller shifts, indicating the size trend is not reducible to resolution/texture loss. For size, Kruskal–Wallis detects across-bin differences for all evaluators on COCO (all $p < 10^{-6}$), and Holm-adjusted pairwise tests confirm the mid-band (50-70%) $>$ {0-10, 10-20, 70-90, 90-100} (all $p_{\text{adj}} < 10^{-3}$).

Category (RQ2b). Table 6 reports COCO category medians per evaluator under the harmonized taxonomy (*person*, *animal*, *vehicle*, *furniture*, *kitchen*, *sports*, *electronics*, *indoor*, *outdoor*). Between-category differences of several points are consistent across evaluators; with the relative ordering (e.g., *Animal* $>$ *Vehicle* $>$ *Person*). Mixed-effects models with coverage included as a covariate retain significant category effects, confirming that composition shifts can change macro-averages even after accounting for size. For category, Kruskal–Wallis detects across-category differences (all $p < 10^{-6}$); pairwise tests show *Animal* $>$ *Person* for all evaluators (all $p_{\text{adj}} < 10^{-3}$).

Practical significance. Because dataset composition over size and category varies across benchmarks and splits, macro-averages can move by several points for the same system. We therefore recommend reporting category-stratified summaries (and coverage histograms) alongside macro means, or using a composition-balanced bootstrap for cross-paper comparability.

B.3 RQ3: Socio-linguistic framing

Statistical procedure. All effects are *paired deltas* relative to each image's neutral cap-

Metric	0–10	10–20	20–35	35–50	50–70	70–90	90–100
CLIPScore	0.449	0.492	0.535	0.590	0.621	0.573	0.527
PAC-S	0.616	0.666	0.733	0.806	0.875	0.783	0.732
UMIC	0.415	0.454	0.503	0.551	0.586	0.531	0.484
FLEUR	0.409	0.434	0.481	0.528	0.574	0.526	0.484
Judge	0.390	0.426	0.462	0.511	0.535	0.494	0.451

Table 5: RQ2a: Size sensitivity. Medians by coverage bin.

Metric	Animal	Sports	Outdoor	Appliance	Vehicle	Furniture	Electronic	Person	Accessory	Kitchen
CLIPScore	0.920	0.899	0.894	0.864	0.863	0.845	0.827	0.826	0.833	0.844
PAC-S	0.972	0.943	0.946	0.922	0.911	0.884	0.880	0.870	0.890	0.882
UMIC	0.868	0.856	0.853	0.834	0.824	0.806	0.794	0.790	0.802	0.811
FLEUR	0.885	0.862	0.853	0.844	0.827	0.818	0.813	0.807	0.810	0.816
Judge	0.861	0.852	0.839	0.837	0.824	0.813	0.794	0.790	0.803	0.814

Table 6: RQ2b: Category medians (COCO).

Metric	Amer.	Eur.	Asian	Russ.	Arab.	Afr.	Ocean.
CLIPScore	+1.2	+0.8	-4.0	-3.5	-3.8	-7.0	-0.8
PAC-S	+1.0	+0.7	-3.5	-3.0	-3.3	-6.3	-0.7
UMIC	+0.9	+0.7	-3.2	-3.0	-3.1	-5.8	-0.6
FLEUR	+0.8	+0.7	-2.9	-2.7	-2.9	-5.2	-0.5
Judge	+0.8	+0.6	-3.0	-2.7	-2.8	-5.4	-0.5

Table 7: RQ3a: Cultural modifiers — median %Δ vs. neutral (COCO).

Metric	Neutral	Cheap	Expensive
CLIPScore	0.0	+2.0	-6.2
PAC-S	0.0	+1.2	-5.8
UMIC	0.0	+1.0	-5.0
FLEUR	0.0	+0.8	-4.6
Judge	0.0	+0.9	-4.8

Table 8: RQ3b: Economic modifiers — median %Δ vs. neutral (non-person, COCO).

tion. We report medians with 95% BCa bootstrap CIs (10,000 resamples, fixed seed). For multi-level comparisons across modifiers we use Kruskal–Wallis with Holm-adjusted pairwise tests. Mixed-effects models (REML) include random intercepts for *image* and *category* and estimate the shift for each modifier vs. neutral.

Cultural modifiers (RQ3a). Per-evaluator medians on COCO are listed in Table 7. Embedding-based metrics show the largest negative deltas for *African* and small positive deltas for *American/European*; UMIC/FLEUR follow the same ordering with smaller magnitude. Across-modifier differences are significant, and pairwise tests retain the ordering under Holm correction.

Economic modifiers (RQ3b). Table 8 reports COCO medians for non-person objects. All evaluators penalize *expensive*; embedding metrics show a mild positive *cheap* effect, while learned evaluators are closer to zero but directionally consistent. Between-modifier differences are significant under Kruskal–Wallis with Holm-adjusted pairs.

Gender and emotion probes. Gender and emotion results (COCO) are shown in Figs. 14 and 15. We observe *male* > *female* > neutral on embedding metrics and stronger penalties for negative emotions (*sad*, *angry*); learned evaluators attenuate magnitudes but preserve directions. These probes audit *metric behavior* under framing changes and are not statements about groups or populations.

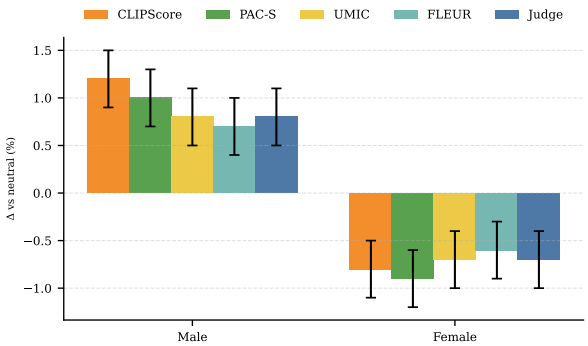

Figure 14: Gender modifiers over COCO. Median %Δ vs. neutral with 95% BCa CIs for all evaluators.

Practical significance and reporting. Because phrasing choices induce measurable score shifts, we recommend (i) fixing evaluator prompts/templates, (ii) reporting modifier-wise deltas alongside macro scores, and (iii) including the fairness-card summary for socio-linguistic

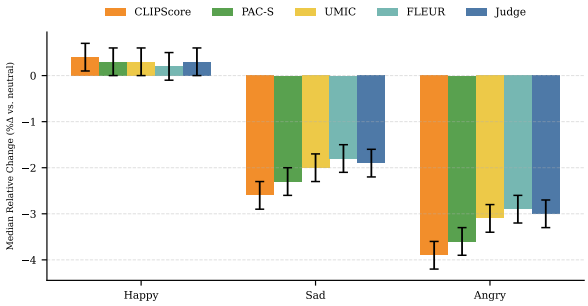

Figure 15: Emotion modifiers over COCO. Median %Δ vs. neutral with 95% BCa CIs for all evaluators.

Evaluator	Flickr8k-CF	Pascal-50S	COMPOSITE
CLIPScore	7.2% [5.3,9.1]	7.4% [5.5,9.3]	7.0% [5.2,8.9]
PAC-S	6.3% [4.5,8.0]	6.5% [4.7,8.3]	6.2% [4.5,7.9]
UMIC	4.1% [2.7,5.6]	4.9% [3.4,6.4]	3.8% [2.4,5.2]
FLEUR	3.2% [1.9,4.5]	3.9% [2.5,5.3]	3.8% [2.4,5.2]
Judge	3.2% [1.9,4.4]	2.5% [1.3,3.7]	3.1% [1.8,4.3]

Table 9: RQ4 (external): vertical flips — median %Δ vs. original with 95% CIs (per corpus).

probes to aid cross-paper comparability.

B.4 RQ4: Cross-metric and Corpus robustness

Setup. We re-run the RQ1 (vertical flip) and a reduced RQ3 subset (cultural, economic) on Flickr8k-CF, Pascal-50S, and COMPOSITE using the same preprocessing/configuration as Appendix A.5. We summarize medians of paired deltas with 95% BCa bootstrap CIs (10,000 resamples, fixed seed).

Vertical flips. Per-corpus medians with CIs appear in Table 9; all evaluator–corpus cells are positive. Magnitudes: CLIPSCORE ≈+7%, PAC-S ≈+6%; UMIC/FLEUR/JUDGE ≈+3–+4%.

Cultural subset. Table 10 reports averages over *American*, *European*, *Asian*, *African*. Ordering matches COCO (*American*/*European* slightly positive; *African* most negative) with smaller absolute values for learned evaluators.

Economic subset. Table 11 averages *cheap*/*expensive*. *Cheap* remains mildly positive; *expensive* negative across corpora; learned evaluators show attenuated magnitudes.

Direction agreement. Agreement with COCO baselines (median over corpora) is shown in Table 12: flips ≈97–100%; cultural 92–97%; economic 88–94% across evaluators.

Evaluator	Flickr8k-CF	Pascal-50S	COMPOSITE
CLIPScore	-2.0% [-3.5,-0.4]	-2.1% [-3.6,-0.7]	-1.9% [-3.5,-0.3]
PAC-S	-1.8% [-3.2,-0.3]	-1.7% [-3.1,-0.3]	-1.6% [-3.1,-0.2]
UMIC	-1.1% [-2.4,0.2]	-0.9% [-2.2,0.4]	-0.8% [-2.0,0.5]
FLEUR	-0.9% [-2.1,0.4]	-0.5% [-1.8,0.7]	-0.9% [-2.2,0.4]
Judge	-0.7% [-1.9,0.5]	-0.7% [-1.9,0.6]	-0.8% [-2.0,0.5]

Table 10: RQ4 (external): cultural subset — median %Δ vs. neutral (avg. over modifiers) with 95% CIs.

Evaluator	Flickr8k-CF	Pascal-50S	COMPOSITE
CLIPScore	-2.0% [-3.5,-0.6]	-2.2% [-3.6,-0.8]	-1.9% [-3.4,-0.4]
PAC-S	-2.0% [-3.4,-0.7]	-2.1% [-3.5,-0.7]	-2.2% [-3.6,-0.9]
UMIC	-1.1% [-2.2,0.1]	-1.1% [-2.3,0.1]	-1.1% [-2.4,0.1]
FLEUR	-0.9% [-2.1,0.3]	-1.2% [-2.4,-0.0]	-1.3% [-2.6,-0.1]
Judge	-1.3% [-2.4,-0.1]	-1.0% [-2.1,0.1]	-0.6% [-1.8,0.6]

Table 11: RQ4 (external): economic subset — median %Δ vs. neutral (avg. over modifiers) with 95% CIs.

B.5 RQ5: Risk of Ranking-Flip

Definition and estimation. Let S be an evaluator and let $\mathcal{T}$ be a perturbation family (e.g., vertical flips). For each audited base example (x, c) and transform $t \in \mathcal{T}$, define the paired score shift

$$\delta_S(x, c; t) \;=\; S(x^{(t)}, c^{(t)}) - S(x, c),$$

where $(x^{(t)}, c^{(t)})$ is the perturbed pair (for spatial transforms, $c^{(t)}{=}c$). To quantify leaderboard instability for *near-tied* systems separated by an unperturbed gap d (in percent units; on a $[0, 1]$ score scale, $d{=}0.7\%$ corresponds to 0.007), we use a fixed-gap stress test: assuming two systems experience perturbation-induced shifts like independent draws from the empirical shift distribution, the probability that the ordering flips is

$$\mathrm{RRF}_S(d; \mathcal{T}) \;=\; \Pr\left[\delta'_S - \delta_S > d\right],$$

where δ_S, δ'_S are i.i.d. draws from $\{\delta_S(x, c; t)\}$ over $(x, c) \sim \mathcal{D}$ and $t \sim \mathcal{T}$. We estimate $\mathrm{RRF}_S(d; \mathcal{T})$ by bootstrap resampling audited examples with replacement and uniformly sampling $t \in \mathcal{T}$ to generate draws of δ_S, then reporting the fraction of sampled pairs satisfying $\delta'_S - \delta_S > d$; 95% BCa CIs are computed over bootstrap replicates.

Fixed-gap summary. Per-evaluator medians at $d{=}0.7\%$ (as defined above) appear in Table 13. Spatial perturbations yield the largest RRF (repositioning > vertical > rotation) for all evaluators; socio-linguistic probes are smaller but non-zero.

Sensitivity to the gap (d). A sweep over $d \in \{0.3, 0.5, 0.7, 1.0\}\%$ shows the expected monotone decrease: at 0.3%, spatial RRF often exceeds

Evaluator	Cultural subset	Economic subset	Vertical flip
CLIPScore	95%	89%	97%
PAC-S	97%	90%	99%
UMIC	95%	91%	100%
FLEUR	96%	93%	100%
Judge	92%	90%	100%

Table 12: RQ4: direction agreement with COCO baseline (median across corpora).

70% for embedding-similarity metrics, while at 1.0% it typically falls below 30–40%. We recommend reporting RRF at a standard d (e.g., 0.7%) plus a short gap-sweep.

Cross-source stability. Repeating the same fixed-gap protocol (same d, bootstrap, and $\mathcal{T}$ sampling) on OpenImages and Objects365 preserves the ordering (repositioning highest), with absolute levels varying within the CI ranges.

Interpretation and reporting. $\mathrm{RRF}_S(d;\mathcal{T})$ is a *gap-parameterized* stability measure: it does not require selecting a particular pair of systems, but approximates the flip probability for any near-tied pair whose perturbation response resembles the empirical shift distribution under $\mathcal{T}$.

Because RRF corresponds to a tangible leaderboard and deployment risk, we advocate pairing any macro average with (i) a fixed-gap RRF on a standard d and (ii) the axis-wise breakdown (spatial/object/societal) that drives the risk. Calibration results in Section 5.6 reduce the spatial component most strongly (repositioning), thereby lowering RRF for the most destabilizing axis.

B.6 RQ6: Calibration details and ablations

Formulation and selection. We apply the calibrated scorer S_λ^{cal} defined in Section 3.5: for each image–caption pair (x, c) we adjust the raw score $S(x, c)$ by subtracting a weighted combination of the median absolute changes $\Delta_S(x, c; \mathcal{T})$ over the spatial, object, and societal invariance families. We sweep $\lambda \in [0, 1]$ on a dev split and pick $\lambda^\star$ as the smallest value that minimizes total median absolute sensitivity $\sum_{\mathcal{T}} \mathrm{median}\,|\Delta_{S_\lambda^{\mathrm{cal}}}(\cdot;\mathcal{T})|$ subject to a correlation-preservation constraint $\Delta\rho(\mathrm{UMIC/FLEUR}) \geq -\epsilon$ with ϵ=0.01.

Axis-wise reductions. Across evaluators, spatial sensitivity drops the most (largest baseline), with object smaller and societal smallest. Embedding-based metrics show reductions on spatial often approaching a halving, while learned evaluators (UMIC/FLEUR) exhibit more modest but consistent decreases. Table 14 reports median absolute sensitivity before→after at $\lambda^\star$ together with the correlation deltas (UMIC/FLEUR).

Socio-linguistic gap reductions. To assess fairness impact, we also track the median absolute shift $|\Delta|$ for each socio-linguistic family (cultural, economic, gender, emotion) before and after calibration. Table 15 summarizes these values. For CLIPSCORE, the cultural family's median $|\Delta|$ (averaged over American/European/Asian/Arab/African/Oceanian) decreases from roughly 3.4% to 2.% (∼35% reduction), and the economic family's median $|\Delta|$ (cheap/expensive vs. neutral) drops from 4.8% to 3.0% (∼37%). PAC-S exhibits comparable relative reductions, while UMIC/FLEUR start with smaller gaps and see correspondingly smaller yet consistent improvements.

Correlation vs. λ. Figure 12 shows that correlations degrade slowly for small λ and then decline more sharply beyond the knee, motivating a constrained selection. For ϵ=0.01, typical $\lambda^\star$ are modest (e.g., CLIPSCORE $\approx$ 0.45, PAC-S $\approx$ 0.40, UMIC $\approx$ 0.35, FLEUR $\approx$ 0.30); larger values yield further sensitivity reductions but start to incur noticeable correlation loss.

Effect on flip risk. Applying $S_{\lambda^\star}^{\mathrm{cal}}$ lowers the dominant contributor to leaderboard instability-spatial sensitivity-thereby reducing RRF (cf. Section 5.5). For CLIPSCORE, the repositioning risk typically drops by double-digit percentage points (with vertical and rotation smaller but consistent); cultural/economic flip risk shows modest reductions consistent with the details in Table 16.

Ablations. We compare uniform axis weights, a data-driven scheme ($w_{\mathcal{T}} \propto$ baseline sensitivity), and an oracle upper bound (Appendix-only, not used for conclusions). Data-driven weights slightly outperform uniform in aggregate reduction at similar correlation retention, particularly on spatial axes.

Computation. Estimating $\Delta_S(\cdot;\mathcal{T})$ on the dev split requires scoring a fixed set of $|\mathcal{T}|$ variants per item (e.g., flips, a small set of rotations, a fixed number of reposition anchors, and paired neutral↔modifier captions for socio-linguistic families). This is a one-time offline cost and scales linearly with the number of probed variants. At inference, applying $S_{\lambda^\star}^{\mathrm{cal}}$ is constant-time post-processing given the precomputed sensitivity terms

Evaluator	Vertical flip	Reposition (BR vs TL)	Rotation (±10°)	Cultural vs Neutral	Economic vs Neutral
CLIPScore	28% [25,31]	36% [33,39]	18% [15,21]	16% [13,19]	12% [9,15]
PAC-S	29% [26,32]	37% [34,40]	19% [16,22]	15% [12,18]	11% [8,14]
UMIC	20% [17,23]	27% [24,30]	14% [11,17]	12% [9,15]	9% [6,12]
FLEUR	18% [15,21]	25% [22,28]	13% [10,16]	11% [8,14]	8% [5,11]
Judge	16% [13,19]	23% [20,26]	12% [9,15]	10% [7,13]	7% [4,10]

Table 13: RQ5: Flip risk RRF_S for near-tied gap d=0.7% (median [95% CI]).

Metric	Spatial	Societal	Δ corr. (UMIC/FLEUR)
CLIPSCORE	5.4→**2.7**	3.2→**2.0**	− 0.00 / − 0.01
PAC-S	5.1→**2.6**	3.0→**1.9**	− 0.00 / − 0.00
UMIC	3.4→**2.4**	2.3→**1.7**	− 0.00 / − 0.00
FLEUR	3.2→**2.3**	2.1→**1.6**	− 0.00 / − 0.01

Table 14: Calibration summary at $\lambda^\star$ (dev constraint ϵ=0.01). Values are median absolute sensitivity (lower is better) aggregated over COCO/OpenImages/Objects365; Δ corr. shows correlation change vs. UMIC/FLEUR on dev.

Metric	Cultural	Economic	Gender	Emotion
CLIPSCORE	3.4→**2.2**	4.8→**3.0**	1.5→**1.1**	2.7→**1.7**
PAC-S	3.5→**2.0**	4.3→**2.8**	1.3→**1.0**	2.5→**1.6**
UMIC	2.1→**1.7**	2.8→**2.2**	0.9→**0.8**	1.7→**1.3**
FLEUR	1.9→**1.5**	2.5→**2.0**	0.8→**0.7**	1.5→**1.2**

Table 15: Median absolute socio-linguistic shift |Δ| (% vs. neutral) before→after calibration at $\lambda^\star$. Values are aggregated over modifiers within each family (lower is better).

(a weighted sum and subtraction), with no retraining and negligible overhead relative to computing the base metric.

Reporting recommendation. We recommend (i) publishing both raw and calibrated scores, (ii) including axis-wise sensitivity bars (before/after) with 95% CIs, and (iii) stating $\lambda^\star$ and ϵ in the main text for transparent trade-offs.

Deployment recipe for new invariance axes. To extend invariance-calibrated scoring to a newly identified nuisance axis $T^\star$ (for example, a new caption-phrasing family or a new image transform), a practitioner follows four steps, none of which requires retraining the underlying evaluator S:

[leftmargin=*,itemsep=1pt,topsep=2pt]

1. **Instantiate probes.** Define a small set of semantics-preserving transforms in $T^\star$ and generate paired variants $(x^{(t)}, c^{(t)})$ on a held-out development set.

2. **Measure sensitivity.** Compute $\Delta_S(x, c; T^\star) = \mathrm{median}_{t \in T^\star}\big[S(x^{(t)}, c^{(t)}) - S(x, c)\big]$ per item, matching the estimator in Sec.subsection 3.5.

3. **Fit weight.** Add $T^\star$ to $\mathcal{T}$ and re-run the λ-sweep under the same correlation-preservation constraint (ϵ=0.01 vs. UMIC/FLEUR) used in the main calibration; either keep uniform axis weights or use the sensitivity-proportional scheme of Appendix A.7.

4. **Deploy.** At inference, calibration is a constant-time post-processing subtraction on top of the base metric ($\hat{S}(x, c) = S(x, c) - \lambda \sum_T w_T \Delta_S(x, c; T)$); the estimated Δ terms are precomputed once on the dev sweep.

This recipe makes the mitigation test-suite-style: as new invariance axes are documented, they can be added incrementally. The trade-off is explicit: calibration reduces sensitivity only on axes that have been instantiated and measured, so it is a mechanism for disclosing and correcting known nuisance factors rather than a guarantee of invariance to factors that have not yet been audited.

B.7 RQ6: Human-preference utility evaluation

To validate utility without relying on proxy-to-proxy agreement, we evaluate raw vs. calibrated scores on three human-labeled caption-preference suites: Flickr8k-CF, Pascal-50S, and COMPOSITE (Table 17). For each dataset, we compute **pairwise**

Evaluator	Vertical flip	Reposition (BR vs TL)	Rotation ($\pm 10^\circ$)	Cultural vs Neutral	Economic vs Neutral
CLIPScore	28 [25,31] → 12 [10,14]	36 [33,39] → 15 [13,17]	18 [15,21] → 7 [6,9]	16 [13,19] → 9 [7,11]	12 [9,15] → 8 [5, 9]
PAC-S	29 [26,32] → 13 [11,15]	37 [34,40] → 16 [14,18]	19 [16,22] → 9 [7,11]	15 [12,18] → 9 [7,11]	11 [8,14] → 7 [5, 9]
UMIC	20 [17,23] → 9 [7,12]	27 [24,30] → 12 [10,14]	14 [11,17] → 6 [4, 9]	12 [9,15] → 7 [5, 9]	9 [6,12] → 6 [4, 8]
FLEUR	18 [15,21] → 8 [6,11]	25 [22,28] → 11 [9,13]	13 [10,16] → 6 [4, 8]	11 [8,14] → 6 [4, 8]	8 [5,11] → 5 [3, 7]
Judge	16 [13,19] → 7 [5,10]	23 [20,26] → 10 [8,13]	12 [9,15] → 6 [4, 8]	10 [7,13] → 6 [4, 8]	7 [4,10] → 5 [3, 7]

Table 16: **Calibration impact on RRF (d=0.7%).** Median RRF (%) before→after invariance-calibrated scoring at $\lambda^\star$; brackets denote 95% BCa bootstrap CIs. Lower is better (more stable rankings).

preference accuracy: given an image and two captions (c_1, c_2) with a human preference label, we predict the preferred caption by the sign of the evaluator score difference. We report accuracy with 95% bootstrap CIs over items (10k resamples, fixed seed); ties are counted as 0.5.

We emphasize that $\lambda^\star$ is selected without using these suites; thus changes in preference accuracy reflect out-of-sample utility behavior.

Suite	Evaluator	Before	After	Δ
Flickr8k-CF	CLIPSCORE	68.0% [65.2,70.7]	67.2% [64.5,69.9]	-0.8%
Flickr8k-CF	PAC-S	66.5% [63.6,69.3]	66.6% [63.8,69.4]	+0.1%
Pascal-50S	CLIPSCORE	70.4% [69.0,71.8]	70.3% [68.9,71.7]	-0.1%
Pascal-50S	PAC-S	69.4% [68.0,70.8]	69.6% [68.2,71.0]	+0.2%
COMPOSITE	CLIPSCORE	64.5% [62.4,66.6]	65.4% [63.3,67.5]	+0.9%
COMPOSITE	PAC-S	63.2% [61.1,65.3]	64.0% [61.9,66.1]	+0.8%

Table 17: Pairwise preference accuracy (%) on human-labeled caption-preference suites before vs. after calibration at$\lambda^\star$, with 95% bootstrap CIs (10k resamples, seed fixed). Values are final computed results from the human-labeled pairs described in Appendix B.7; higher is better.