A Diagnostic Framework for Auditing Reference-Free Vision-Language Metrics

Angeline Charles[1,*,†], Srikant Panda[2,*,†], Amit Agarwal[2,†], Hitesh Laxmichand Patel[2,†], Priyaranjan Pattnayak[2,†], Bhargava Kumar[3,†], Tejaswini Kumar[4]

[1]Moody's Analytics, [2]Oracle AI, [3]TD Securities, [4]Columbia University

Correspondence: srikant86.panda@gmail.com

[*]Equal contribution

Abstract

Reference-free metrics such as CLIPScore and PAC-S are increasingly used in vision-language tasks due to their scalability and independence from human-written references. However, their reliability under linguistic, visual, and cultural variation remains underexplored. In this work, we present a systematic audit of CLIPScore and PAC-S using an eight-factor diagnostic framework applied to MS-COCO validation images. Our analysis reveals consistent failure modes across dimensions including object size, content category, syntax, named entities, spatial relations and cultural context. Both metrics penalize captions referencing *African* (-5.5%, -4.8%) and *Arabian* (-4.9%, -5.3%) cultures, favor large-object and animal-centric scenes (by 20-30%) and show limited sensitivity to spatial negation and word order. CLIPScore correlates more strongly with syntactic complexity, while PAC-S demonstrates greater robustness to verbosity and named–entity variation highlighting complementary strengths rather than superiority. These findings expose cultural and content bias, weak semantic robustness, and limited compositional understanding. We conclude with design recommendations to improve fairness, scale invariance, and semantic grounding in future reference-free evaluation metrics. [1]

1 Introduction

The rise of multimodal large language models (MLLMs) (Liu et al., 2023) has enabled significant advances in vision–language tasks, including image captioning, text-to-image generation, and visual question answering. As these systems generate increasingly fluent and contextually grounded outputs, the need for reliable evaluation becomes more critical. Evaluation metrics play a central role in this ecosystem–they benchmark model performance, shape training objectives, and inform deployment decisions.

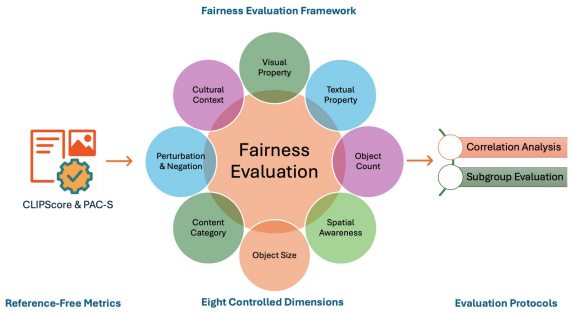

Figure 1: Reference-Free Metrics Evaluation Framework

Historically, reference-based metrics such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016) have dominated image–text evaluation. These metrics compare generated outputs to fixed sets of human-written references and provide interpretable, reproducible scores. However, their reliance on limited references makes them brittle in open-ended generation settings, where linguistic diversity is a feature rather than a flaw. They frequently penalize factually correct yet stylistically novel captions, limiting their usefulness for flexible or creative generation.

To address these limitations, reference-free metrics such as **CLIPScore** (Hessel et al., 2021) and **PAC-S** (Sarto et al., 2023a) have gained prominence for their scalability and independence from reference captions. These metrics leverage pretrained vision–language models to assess semantic alignment without requiring ground-truth captions. However, despite their scalability, their reliability across linguistic, visual, and cultural variations remains poorly understood.

In this work, we present a systematic audit of CLIPScore and PAC-S using a controlled diagnostic framework. Treating these metrics themselves as systems under test, we evaluate their behavior across eight dimensions including syntax, object

[1]†Work done outside of primary affiliation

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2633–2644

size, spatial relations and cultural context. Our analysis is based on 5,000 curated MS-COCO validation images, enabling reproducible and fine-grained evaluation.

Our contributions are as follows:

- We propose a diagnostic framework for auditing reference-free metrics across linguistic, visual, and cultural axes.
- We uncover consistent biases and failure modes in CLIPScore and PAC-S, including cultural bias, scale sensitivity and limited semantic robustness.
- We provide actionable design recommendations to improve fairness, compositional understanding, and semantic grounding in future metrics.

These findings highlight the need for more equitable and interpretable evaluation tools as multimodal systems continue to evolve.

2 Related Work

Emergence of reference-free evaluation metrics has significantly reshaped the landscape of vision-language evaluation. Among the most widely adopted is CLIPScore, which estimates image-text similarity using CLIP embeddings and has demonstrated superior performance over reference based metrics. Železný (2023) established its robustness across MS-COCO, while Cho et al. (2023a) employed CLIPScore to reward semantic specificity during caption generation. Barraco et al. (2022) further solidified CLIP's role as a powerful visual encoder, helping establish CLIPScore as a de facto semantic metric. This established utility, however, leaves open questions around its sensitivity to spatial structure, compositionality, & cultural nuance.

To address some of these limitations, contrastive learning-based metrics have gained traction. PAC-S, proposed by Sarto et al. (2023a), employs augmented-positive contrastive learning to improve alignment with human preferences and detect hallucinations more effectively. Its successor, PAC-S++, offers improved sensitivity to syntactic noise and redundant phrasing. Complementary approaches such as HICE-S (Zeng et al., 2024) and comparative analyses by González-Chávez et al. (2023) underscore the growing interest in contrastive and multi-scale evaluation strategies. As a result, PAC-S represents a valuable counterpoint to CLIPScore in our comparative analysis.

Building on core paradigms, several recent works have explored architectural strategies for improving evaluation reliability. Fusion-based methods such as ECO (Jeong et al., 2024) & BRIDGE (Sarto et al., 2023b) aggregate multiple metric signals to improve caption ranking and hallucination detection. Ross et al. (2024) argue that current T2I metrics over-rely on surface-level textual overlap, while Wu et al. (2018), through their work on visual change detection, highlight the challenge of evaluating object relationships and spatial directionality challenges we explore through prompt perturbation. These innovations inform our methodological choice to apply structural interventions & test metrics.

Parallel to architectural advances, optimization-based efforts have focused on tuning metric behavior. ReCap, by Paischer et al. (2025), demonstrates that fine-tuning alignment layers can enhance semantic fidelity in vision-language models, while Kornblith et al. (2023) show that classifier-free guidance can yield more expressive and stable generations, highlighting the importance of embedding calibration in metric performance. These insights guide our use of controlled test conditions to isolate metric behavior under shared embeddings.

While these developments have advanced the field, a growing body of work has drawn attention to the limitations and blind spots of reference-free metrics. Ahmadi and Agrawal (2024) and Kasai et al. (2022) question whether popular metrics like CLIPScore and PAC-S adequately reflect human preferences or linguistic complexity. Zur et al. (2024) surface accessibility concerns, especially for blind and low-vision users, showing that CLIP-based metrics poorly assess utility-driven captioning. In response, Lee et al. (2024) propose FLEUR, a rationale-aligned and explainable evaluation framework. These critiques underscore the importance of interrogating biases, fairness, and cultural representation in metric behavior dimensions that we place at the center of our analysis.

Together, these contributions form the foundation for our study. They reveal that while metrics like CLIPScore and PAC-S perform well on average correlation benchmarks, they may fail under structured stressors, cultural shifts, or compositional transformations. Our work builds on these insights by systematically auditing these metrics across multiple controlled axes such as object count, syntax, spatial relations, and cultural cues using MS-COCO as a testbed for fine-grained diagnostic evaluation.

3 Methodology

3.1 Diagnostic Framework

To critically assess the reliability and fairness of reference-free evaluation metrics, we propose a systematic diagnostic framework that treats the metrics themselves as *systems under test*. Rather than assuming these metrics to be reliable surrogates for human judgment, we audit their behavior across a diverse set of diagnostic axes designed to reveal hidden biases, robustness gaps, and semantic insensitivities. Our analysis focuses on reference-free metrics: CLIPScore and PAC-S. We restrict our evaluation to these metrics not only because they are prominent in current multimodal evaluation, but also because they are most compatible with a controlled diagnostic setup that isolates metric behavior. Both methods rely on shared CLIP-based embeddings, ensuring comparability under identical input conditions.

We exclude other methods such as *UMIC*(Lee et al., 2021), *TIFA*(Hu et al., 2023), *VPEval*(Cho et al., 2023b) and *DSG*(Cho et al., 2024) for specific methodological reasons. UMIC depends on the UNITER(Chen et al., 2020) architecture, which processes images and text through decoupled pipelines, complicating the attribution of evaluation behavior and reducing transparency, contrary to our goal of treating metrics as interpretable systems. Metrics like TIFA, VPEval and DSG are VQA-based, requiring multiple grounded questions per image; however, our single-dominant-object setup permits only one meaningful grounded query, limiting their ability to evaluate fine-grained semantic variation. By constraining the study to CLIPScore and PAC-S, we ensure that our audit remains transparent, interpretable, and computationally tractable (≈25,000 evaluations). This design enables us to isolate systematic biases (e.g., cultural, content, scale, and syntax) without interference from architecture or dataset specific confounds introduced by more complex evaluators.

We structure our audit around the following guiding question: "*How well do these metrics satisfy the key desiderata of a good evaluator?*" Specifically, we assess:

- **Scene understanding** – Can metrics handle dense, compositional, or complex inputs?
- **Linguistic alignment** – Do metrics reward relevance and precision over verbosity?
- **Fairness** – Are scores invariant to cultural or contextual variation?
- **Semantic sensitivity** – Do metrics reliably distinguish correct from incorrect text?

This framing allows us to evaluate metric behavior across linguistic and visual dimensions while maintaining reproducibility and methodological control, establishing a foundation for the diagnostic analyses presented in subsequent sections.

3.2 Dataset Construction

We use MS-COCO 2017 validation set (Lin et al., 2015) as our evaluation benchmark. It is a widely used Common Objects corpus used in vision-language research, containing richly annotated images with bounding boxes and multiple human-written captions. Images were filtered to contain single occurance of the objects identified through bounding box size, object count and category labels. Specifically, we selected images in which a single object class appeared exactly once in the scene, ensuring an unambiguous visual target per image and avoiding the semantic ambiguity that arises from multiple instances of the same category. For example, if an image contained a bicycle, a toothbrush, and two spoons, only the unique objects (bicycle and toothbrush) were retained, while categories with repeated instances (spoons) were discarded.

Although the dataset emphasizes single-object clarity, we also incorporated controlled synthetic spatial relations to probe metric sensitivity to spatial language (e.g., "There is a cycle left of a toothbrush" / "There is a toothbrush right of a cycle"). These captions were programmatically generated using fixed templates while maintaining syntactic coherence and visual grounding. This design allowed us to test spatial awareness without violating the single-dominant-object constraint, since relations were applied only between distinct unique categories within the filtered set. To ensure representativeness, the subset was stratified across MS-COCO supercategories and cultural modifiers, enabling broad semantic coverage despite its reduced scale. The decision to limit the sample to 5,000 images was driven by computational feasibility; each image-caption pair was evaluated by two metrics under multiple perturbation axes, amounting to ≈25,000 evaluations and also comparable scales have been adopted in prior studies, including González-Chávez et al. (2023), Wu et al. (2018), and Kasai et al. (2022), which demonstrate that such subsets are sufficient for robust correlation and bias analysis. Captions were drawn from two

complementary sources:

- **Natural captions** – the five human-authored MS-COCO captions per image, preserving natural linguistic variation.
- **Fixed-format captions** – synthetically generated templates (e.g., "There is a/an [object]", "There is a/an [cultural_modifier] [object]", "There is a/an object_i left of object_j" and There is a/an object_j right of object_i"), created to isolate specific linguistic or cultural factors while maintaining semantic consistency.

This hybrid setup balances experimental control and ecological validity, reconciling real-world linguistic diversity with reproducible, fine-grained diagnostics. A full visualization of the dataset construction pipeline and filtering process is provided in **Appendix A**, Figure 8.

3.3 Evaluation Setup

As a sanity check and to establish baseline behavior, we compute score distributions (Figure 2) and conduct statistical comparisons (Table 1) between **CLIPScore** and **PAC-S** over all image-text pairs. Using paired t-tests (Student, 1908) and Pearson correlation (Spearman, 2015), we quantify both agreement and divergence, setting the stage for deeper diagnostic evaluation. The results indicate that CLIPScore and PAC-S exhibit complementary behaviors rather than interchangeable performance. CLIPScore tends to assign lower but more tightly clustered scores, whereas PAC-S displays a broader distribution with higher mean values.

We report both statistical significance and effect sizes, and treat differences as meaningful only when (i) results are statistically significant ($p < 0.05$) and (ii) the effect magnitude exceeds a practical threshold (e.g., $\geq$3–5% deviation for subgroup comparisons). This ensures that small or non-significant deviations are not automatically interpreted as bias, but rather as normal variability.

Distributions of metric scores are visualized in Figure 2, showing that CLIPScore values are more concentrated around the mean, while PAC-S exhibits a wider spread and slightly higher central tendency. The moderate correlation ($r = 0.53$) (Table 1) suggests that the two metrics capture related but distinct aspects of image-text alignment, supporting our decision to audit both jointly across diagnostic axes.

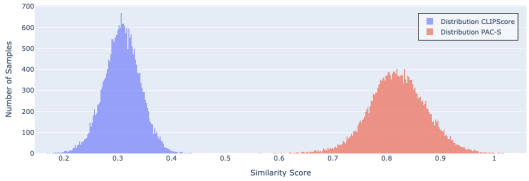

Figure 2: Score distribution comparison between CLIPScore and PAC-S.

Test	Statistic	P-value
T-test	-1266.19	0.0000
Pearson correlation	0.5326	0.0000

Table 1: Statistical comparison between CLIPScore and PAC-S

3.4 Evaluation Protocol

We employ two complementary strategies to analyze metric behavior.

Correlation Analysis: We compute **Spearman** and **Pearson** correlations between metric scores and textual or visual properties (e.g., object count, color variance, caption complexity). Strong correlations in unintended directions indicate potential bias. Only statistically significant results ($p < 0.05$) with practical effect sizes are interpreted as meaningful.

Subgroup Comparison: We compare average scores across controlled subgroups (e.g., "American" vs. "African", "small" vs. "large" objects) to assess fairness and consistency. Differences are considered substantial when exceeding 3-5% and statistically significant.

This dual approach combines quantitative sensitivity analysis with interpretable group-level comparisons, capturing both general trends and localized failure modes. The metrics are evaluated across eight diagnostic dimensions summarized in Table 2.

4 Results

4.1 Textual Property - Evaluating Sensitivity to Linguistic Structure

An ideal evaluation metric should reward captions that accurately describe image content rather than those that are longer or more complex. In this section, we examine whether **CLIPScore** and **PAC-S** are overly sensitive to surface-level language features. To examine this, we analyze the correlation between metric scores and four textual attributes:

Axis	Description	Caption Type	Eval Protocol
Text Properties	Caption length, syntax complexity, passivity	Original	Corr. Analysis
Visual Properties	Entropy, sharpness, color, edge complexity	Original	Corr. Analysis
Object Count	Number of distinct objects in image	Original	Corr. Analysis
Cultural Context	7 Fixed Cultural references	Fixed Format	Subgroup Eval.
Content Category	MSCOCO Category references	Fixed Format	Subgroup Eval.
Object Size	Percent of image area covered by object	Fixed Format	Subgroup Eval.
Spatial Awareness	Absolute and relative object positioning	Fixed Format	Subgroup Eval.
Perturbations	Grayscale, negation, word order changes	Original	Subgroup Eval.

Table 2: Evaluation framework across key diagnostic dimensions for metric auditing.

1. **Text Length**: Total number of non-stopword tokens.
2. **Sentence Complexity**: Ratio of tokens and noun phrases to the number of clauses, approximating syntactic density.
3. **Flesch–Kincaid Grade Level**: Approximate U.S. school grade required to comprehend a caption.
4. **Named Entity Count**: Number of recognized named entities such as people, locations, or organizations.

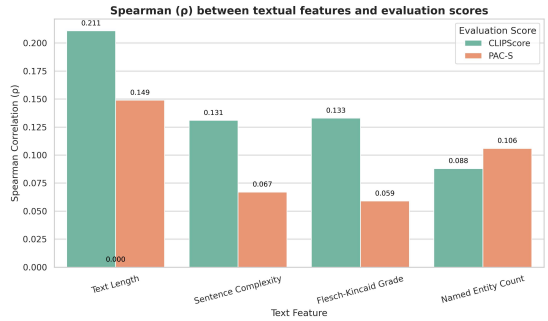

Figure 3: Spearman (ρ) between textual features and evaluation scores.

Observation: As observed in Figure 3, CLIPScore exhibits a notable positive correlation with caption length ($\rho = 0.211$) and syntactic complexity, indicating a tendency to favor longer or more elaborate phrasing. PAC-S also shows a positive but weaker correlation ($\rho = 0.149$) and is less influenced by sentence structure. Additionally, CLIPScore correlates positively with sentence complexity and readability grade, indicating sensitivity to caption structure, while PAC-S appears less affected. Interestingly, PAC-S displays slightly higher responsiveness to the presence of named entities ($\rho = 0.106$), suggesting an implicit bias toward entity-rich descriptions.

Metric Expectation: Metrics should score captions based on semantic relevance and remain invariant to linguistic verbosity and structural variation.

Failure Mode: CLIPScore tends to penalize concise yet accurate captions, while PAC-S favors simpler phrasing but rewards named-entity presence.

4.2 Visual Property-Testing Robustness to Low-Level Image Attributes

Metrics for vision-language evaluation should remain stable under low-level visual variations such as color richness, texture, or structural detail when the semantic meaning of an image is unchanged. We evaluate whether **CLIPScore** and **PAC-S** are affected by these visual features. All images are resized to 224×224 pixels and normalized to [0,1] prior to feature computation. We extract three descriptive visual attributes:

1. **Color Variance:** Measures the average variance across RGB channels-higher values indicate richer color diversity.
2. **Energy and Homogeneity:** Derived from the Gray-Level Co-occurrence Matrix (GLCM), these texture features capture local pixel relationships without altering image semantics.
3. **Edge Density:** Computed using the Canny edge detector as the ratio of edge pixels to total image pixels, indicating visual detail.

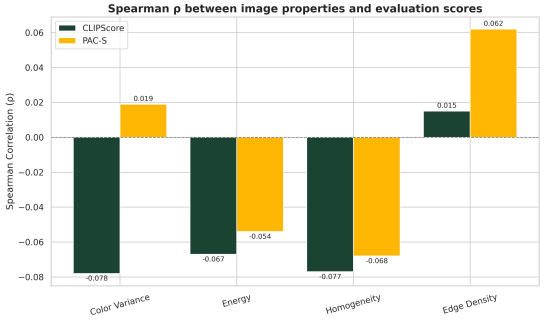

Figure 4: Spearman ρ between image properties and evaluation scores.

Observation: As observed in Figure 4 **CLIPScore** shows a weak negative correlation with color variance ($r = -0.078$), indicating a mild penalty for visually diverse images, while **PAC-S** remains largely unaffected ($r = 0.019$). Both metrics also exhibit weak negative correlations with texture-based features such as energy and homogeneity, suggesting slight penalties for highly textured or overly uniform images irrespective of semantic correctness. Additionally, **PAC-S** shows a slight preference for images with higher edge density ($r = 0.062$), reflecting a modest bias toward more detailed or structured visuals, whereas **CLIPScore** remains mostly invariant.

Metric Expectation: An ideal evaluation metric should remain invariant to low-level visual changes that do not affect semantic alignment.

Failure Mode: While **CLIPScore** slightly penalizes images with richer color variance or texture, **PAC-S** tends to favor sharper, edge-dense visuals. These tendencies reveal that both metrics retain limited robustness to low-level image attributes, coupling semantic judgment with perceptual quality.

4.3 Object Count – Assessing Compositional Generalization

A reliable evaluation metric should effectively handle complex scenes containing multiple distinct objects, as commonly found in real-world environments such as surveillance, robotics, and captioning benchmarks. Captions describing such images should not be penalized merely due to scene complexity. We evaluate whether **CLIPScore** and **PAC-S** are sensitive to object count by correlating their evaluation scores with the number of distinct object classes per image, computed from MS-COCO annotations.

Observation: As observed in Table 3 Both metrics exhibit a small but consistent negative correlation with object count, suggesting that evaluation scores tend to decrease as the number of objects increases. This pattern indicates that both metrics slightly undervalue captions describing more complex scenes, potentially due to difficulty in grounding multiple entities simultaneously.

Metric Expectation: Metrics should evaluate captions based solely on their semantic accuracy and grounding, irrespective of how many objects are present.

Feature	CLIPScore	PAC-S
Object Count	-0.084	-0.080

Table 3: Spearman (ρ) between object count and evaluation scores for both metrics

Failure Mode: Both **CLIPScore** and **PAC-S** demonstrate mild sensitivity to scene complexity, penalizing captions in multi-object settings. This reveals a limited capacity for compositional generalization, as both metrics struggle to maintain consistent alignment when multiple visual entities interact within a single frame.

4.4 Cultural Context – Auditing Cultural Fairness in Evaluation

A reliable evaluation metric should be culturally agnostic, giving comparable scores to semantically identical captions regardless of geographic or cultural modifiers. To test this, we evaluated how **CLIPScore** and **PAC-S** respond to captions that differ only by cultural adjectives, while maintaining identical syntax and object identity. For each single-object image (e.g., "chair," "car"), we created fixed-syntax captions of the form: *"There is a/an [American / African / Asian / European / Russian / Arabian / Oceania] [object name]."*. The image remained constant across all variants, allowing us to isolate the effect of the cultural term itself. Average scores across regions are shown in Figure 5.

Observation: Both metrics consistently assign lower scores to culturally modified captions compared to the neutral form, revealing a uniform drop across modifiers. **CLIPScore** shows the strongest bias against African (-5.5%) and Arabian (-4.9%) descriptors, while **PAC-S** registers similar declines for Arabian (-5.3%) and Oceania (-5.2%) terms. American and European modifiers receive scores closest to the baseline, indicating a clear Western-centric bias in both evaluation measures.

Metric Expectation: Scores should depend solely on the semantic correctness and grounding of the caption.

Failure Mode: Both metrics demonstrate systematic Western preference, lowering scores for non-Western cultural adjectives even under syntactically fixed and semantically equivalent conditions. This pattern suggests that pretraining data distribu-

tions and embedding representations contribute to inherited cultural bias.

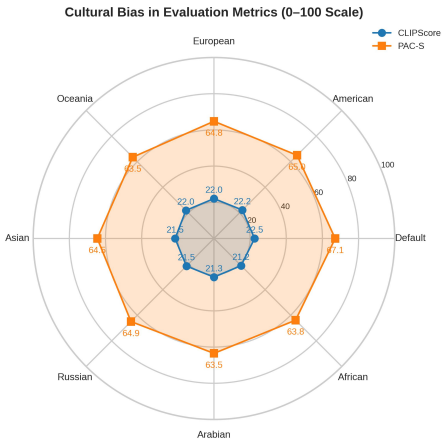

Figure 5: Cultural bias analysis using a radar plot showing evaluation scores (on a 0-100 scale) across various cultural regions.

4.5 Object Category – Evaluating Content-Type Sensitivity

A fair evaluation metric should assess captions based on semantic correctness, irrespective of the type of visual content whether it depicts animals, humans, objects, or environmental elements. Systematic variation in scores across content types, without semantic justification, indicates domain-level bias. To evaluate this, we analyzed average metric scores across 12 MS-COCO supercategories, using images containing a single dominant object from each category. Fixed-format captions ("There is a/an object_name") were used to ensure consistency in linguistic structure. Figure 6 presents the mean CLIPScore and PAC-S results for all categories.

Observation: Both metrics exhibit consistent content-type bias. Animal-related images receive the highest scores (CLIPScore: 0.2508; PAC-S: 0.7341), reflecting an overemphasis on easily recognizable subjects. Appliance and sports scenes also score relatively high, whereas person, kitchen, and accessory categories show the lowest average scores. The person category demonstrates the strongest negative deviation, 16.2% below the mean for CLIPScore and 11.6% for PAC-S indicating that both metrics systematically undervalue human-centric or indoor scenes.

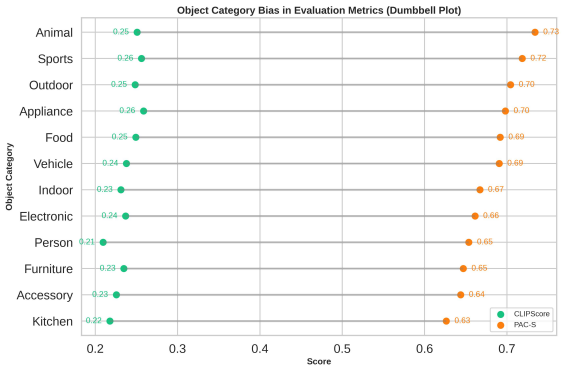

Figure 6: Dumbbell plot showing Object Category bias, indicating metric sensitivity to semantic content.

Metric Expectation: An ideal metric should provide consistent evaluations across object types when captions are semantically accurate, avoiding preferential treatment for specific visual domains.

Failure Mode Both CLIPScore and PAC-S reveal domain bias, favoring animal, sports, and appliance scenes while penalizing person-centric and indoor content. These trends likely stem from training data imbalances in CLIP and related embedding models.

4.6 Object Size – Evaluating Scale Sensitivity and Visual Prominence Bias

An effective evaluation metric should be scale-invariant, assigning similar scores to accurate captions regardless of object size or prominence. Otherwise, it risks undervaluing small-object recognition or penalizing captions for cluttered or zoomed-out scenes. To examine scale sensitivity, we grouped images by object-area percentage, the proportion of the image occupied by the single-dominant-objects, and computed average metric scores using fixed-format captions ("There is a/an object_name") to isolate the effect of size while keeping caption syntax constant. Figure 7 presents average scores across size bins.

Observation: Evaluation scores increase with object size, revealing a clear dependence on scale. Both CLIPScore and PAC-S peak in the 60-80% range, favoring medium-to-large, clearly visible objects. However, performance drops at both extremes: very small objects (0-10%) receive lower scores, likely due to difficulty in grounding fine details, while extremely large objects (90-100%) also score lower, possibly from loss of contextual grounding in cropped or zoomed-in views.

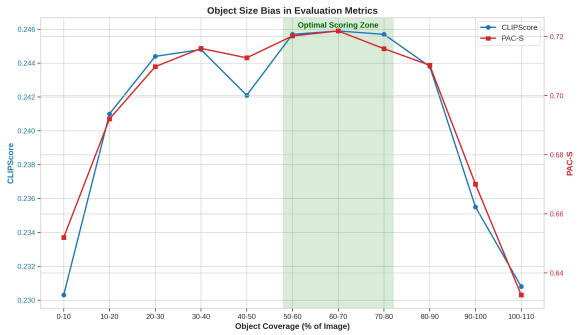

Figure 7: Evaluation metrics (CLIP and PAC-S) peak within the 50-80% object coverage range, indicating a bias toward medium-sized objects.

Metric Expectation: Metric should maintain consistent scoring across object scales, reflecting semantic correctness rather than visual prominence.

Failure Mode: Both metrics favor mid-sized objects and undervalue very small or very large ones (90-100% of the image area), limiting their reliability in tasks requiring fine-grained or contextual grounding. This trend likely reflects the data distribution bias of pretrained embedding models.

4.7 Spatial Awareness – Testing Positional Sensitivity and Object Relations

Spatial awareness is a crucial component of visual-linguistic understanding, enabling models and metrics to reason about object placement and relationships within an image. A reliable evaluation metric should assign consistent scores to captions describing the same spatial configuration, regardless of object orientation or phrasing. To evaluate this capability, we analyze whether CLIPScore and PAC-S remain invariant under changes in absolute object position and relative spatial relations between distinct objects.

Table 4: Mean scores for absolute vs. relative positioning; % differences are relative to baseline. * indicates baseline, Abs.-Absolute, Rel.-Relative, L-Left and R-Right)

Positioning Type	**CLIPScore**	**PAC-S**
Abs.: Left*	0.2281	0.6805
Abs.: Right	0.2281	0.6803
	(0.0%)	**(-0.02%)**
Rel.: L to R*	0.2301	0.667
Rel.: R to L	0.2337	0.6620
	(+1.5%)	**(-0.07%)**

Experimental Setup:

1. **Absolute Positioning:** To test positional bias, we compare metric scores for identical captions describing objects located on different sides of the image. Left and right configurations are created by horizontal image flipping, ensuring that only the object's position changes while the caption remains fixed (There is a/an 'object_name').
2. **Relative Positioning:** To test relational bias, we generate semantically equivalent but syntactically reversed captions for pairs of distinct objects such as: i) There is a/an [object_i] left to [object_j] and ii) There is a/an [object_j] right to [object_i].

Observation: As observed in Table 4 Both metrics show strong invariance to absolute positioning, producing nearly identical scores for left and right object placements. However, in relative positioning, CLIPScore exhibits a minor asymmetry, slightly favoring "right of" over "left of" phrasing (+1.5%), suggesting sensitivity to linguistic formulation rather than visual semantics. PAC-S, by contrast, remains largely stable across both cases.

Metric Expectation: Metrics should treat equivalent spatial relationships equally, regardless of orientation or caption order, as both convey identical semantic meaning.

Failure Mode: While PAC-S demonstrates stable spatial awareness, CLIPScore reveals directional phrasing bias, indicating mild sensitivity to syntactic ordering in relative spatial expressions. This asymmetry implies partial reliance on textual embeddings over grounded spatial understanding.

4.8 Perturbations & Negations

Robust evaluation metrics should accurately distinguish semantically correct captions from incorrect or syntactically degraded ones, while remaining stable under irrelevant visual changes. We assess the resilience of CLIPScore and PAC-S to controlled perturbations in spatial accuracy, visual appearance, and word order. We evaluate the following categories of perturbations,

1. **Relative Spatial Negation:** We switch object positions in captions to create mismatches (e.g., "There is a [object A] left of [object B]" vs. incorrect "right of" when A is actually on the left).

Perturbation Category	Condition	CLIPScore	PAC-S
Absolute Position	Correct placement	0.2356	0.6594
	Incorrect placement	0.2354 **(-0.08 %)**	0.6591 **(-0.04 %)**
Relative Position	Correct referenced	0.2301	0.6670
	Incorrect referenced	0.2340 **(+0.5 %)**	0.6626 **(-0.7 %)**
Multimodal Augmentation	Original image and caption	0.3077	0.8204
	Black & white image	0.2996 **(-2.63 %)**	0.8110 **(-1.15 %)**
	Reverse word order	0.2836 **(-8.70 %)**	0.8015 **(-2.38 %)**
	Random word order	0.2769 **(-10.28 %)**	0.7937 **(-3.29 %)**

Table 5: Evaluation scores for spatial negation and multimodal perturbations.

2. **Absolute Spatial Negation:** We flip spatial terms like "left" and "right" in captions (e.g., "There is a [object A] on the left side" vs. incorrect "right side" when A is on the left).
3. **Multimodal Input Perturbations:** Convert image grayscale (tests visual robustness), Reverse word order (tests mild syntactic distortion) and Randomize word order (tests full syntactic corruption)

Observation: As visible in Table 5, CLIPScore frequently fails to penalize spatially incorrect captions, occasionally assigning slightly higher scores than correct ones. PAC-S performs marginally better but with minimal margin. Both metrics remain mostly unaffected by grayscale conversion, showing visual invariance. However, they exhibit limited sensitivity to syntactic corruption-maintaining relatively high scores even for reversed or randomized captions-revealing bag-of-words behavior that overlooks sentence structure.

Metric Expectation: Metrics should penalize semantically incorrect or negated captions, remain robust to visual changes, and clearly reduce scores when sentence structure loses grammatical or semantic integrity.

Failure Mode: Both CLIPScore and PAC-S demonstrate low semantic robustness, failing to effectively distinguish between correct and negated or disordered captions. Their overreliance on token-level similarity instead of sentence-level meaning limits their ability to capture true compositional semantics.

5 Summary of Metric Behavior

We provide a summary of the diagnostic behavior of CLIPScore and PAC-S on all axes of evaluation in Table 6 in Appendix B. Both provide scalable, reference-free evaluation, but our analysis demonstrates a number of reliable shortcomings: Visual and Textual Bias, Cultural Bias, Content-Type Bias, Scale & Object Count Sensitivity, Spatial Robustness and Perturbation Weakness. Overall, these findings highlight systematic dependence on textual verbosity, cultural framing, and visual prominence, emphasizing the need for future metrics that combine semantic grounding, cultural fairness, and compositional understanding for robust multimodal evaluation.

6 Conclusion

Reference-free metrics like CLIPScore & PAC-S are gaining traction in vision-language research due to their scalability and independence from annotated references. However, our analysis shows they often fail to align with human judgment across diverse contexts.

We identify key limitations, including over reliance on surface features, low robustness to syntactic variation, and cultural biases e.g., consistently lower scores for modifiers like "African" & "Arabian." These findings raise concerns about their equitability and generalizability.

To address these gaps, we recommend: (1) prioritizing semantic grounding over shallow cues; (2) ensuring fairness across cultures, geographies, and object categories; (3) maintaining robustness in complex, multi-entity scenes; (4) penalizing syntactic or factual errors; (5) improving transparency through interpretable diagnostics; and (6) expanding fairness evaluation to underrepresented group.

We hope these guidelines inform the development of reference-free metrics that are equitable, interpretable, and reliable. As multimodal systems advance, robust evaluation standards will be essential to ensure meaningful progress.

7 Limitations

Although our study provides a systematic and transparent audit of reference-free evaluation metrics, it is bounded by several methodological limitations. Our dataset design prioritized experimental control by focusing on single-object images, which enabled clear attribution of metric behavior but limited applicability to real-world, multi-object scenes that involve more complex spatial and compositional reasoning. The dataset size ($\approx$ 5,000 images, $\approx$ 25,000 evaluations) was chosen for computational tractability and consistency with prior studies, yet this smaller scale constrains generalizability to larger or more diverse datasets. In addition, the cultural fairness audit, though broader than in previous work, covered only seven modifiers and corresponding global regions, leaving finer-grained cultural or linguistic variations underexplored. The use of fixed-format captions further ensured semantic control but could not capture the richness and ambiguity of natural human language, which may influence how metrics respond to real-world linguistic diversity.

Beyond dataset factors, our analysis was restricted to two representative metrics, CLIPScore and PAC-S excluding emerging VQA-based and LLM-based scoring methods due to architectural incompatibility with our diagnostic setup. Finally, while our framework combined quantitative and qualitative analyses, it relied primarily on subgroup analysis and correlation-based evaluations that do not fully capture nonlinear or interdependent effects between linguistic and visual attributes. Future work should address these limitations by expanding dataset diversity and scale, incorporating multi-object and context-rich scenes, and extending the framework to a broader range of metrics and nonlinear analytical models to enable more comprehensive and inclusive auditing of vision–language evaluation systems.

References

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. *ArXiv*, abs/1607.08822.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

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.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. *Preprint*, arXiv:1909.11740.

Jaemin Cho, Yushi Hu, Roopal Garg, Peter Anderson, Ranjay Krishna, Jason Baldridge, Mohit Bansal, Jordi Pont-Tuset, and Su Wang. 2024. Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. *Preprint*, arXiv:2310.18235.

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

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023b. Visual programming for text-to-image generation and evaluation. *Preprint*, arXiv:2305.15328.

Othón González-Chávez, Guillermo Ruiz, Daniela Moctezuma, and Tania A. Ramirez-delReal. 2023. Are metrics measuring what they should? an evaluation of image captioning task metrics. *Preprint*, arXiv:2207.01733.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *ArXiv*, abs/2104.08718.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. 2023. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. *Preprint*, arXiv:2303.11897.

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.

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.

Simon Kornblith, Liunian Harold Li, Zhe Wang, and Thien Huu Nguyen. 2023. Classifier-free guidance makes image captioning models more descriptive. In *ICLR 2023 Workshop on Multimodal Representation Learning: Perks and Pitfalls*.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Trung Bui, and Kyomin Jung. 2021. Umic: An unreferenced metric for image captioning via contrastive learning. *ArXiv*, abs/2106.14019.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. Fleur: An explainable reference-free evaluation metric for image captioning using a large multimodal model. *Preprint*, arXiv:2406.06004.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Fabian Paischer, Markus Hofmarcher, Sepp Hochreiter, and Thomas Adler. 2025. Linear alignment of vision-language models for image captioning. *Preprint*, arXiv:2307.05591.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

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.

C. Spearman. 2015. The proof and measurement of association between two things. *International journal of epidemiology*, 39 5:1137–50.

Student. 1908. The probable error of a mean. *Biometrika*, 6:1–25.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. *Preprint*, arXiv:1411.5726.

Junhui Wu, Yun Ye, Yu Chen, and Zhi Weng. 2018. Spot the difference by object detection. *Preprint*, arXiv:1801.01051.

Tomáš Železný. 2023. Exploring the relationship between dataset size and image captioning model performance. Unpublished manuscript.

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.

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 Supplementary Details on Dataset Construction

Figure 8 provides a visual overview of our dataset construction pipeline, illustrating the filtering of MS-COCO images, object class extraction, and the generation of both natural and fixed form captions used in our experiments.

B Qualitative Summary of Metric Behavior

We present in Table 6 a qualitative comparison of CLIPScore and PAC-S across diagnostic axes, highlighting observed biases and deviations from ideal metric behavior.

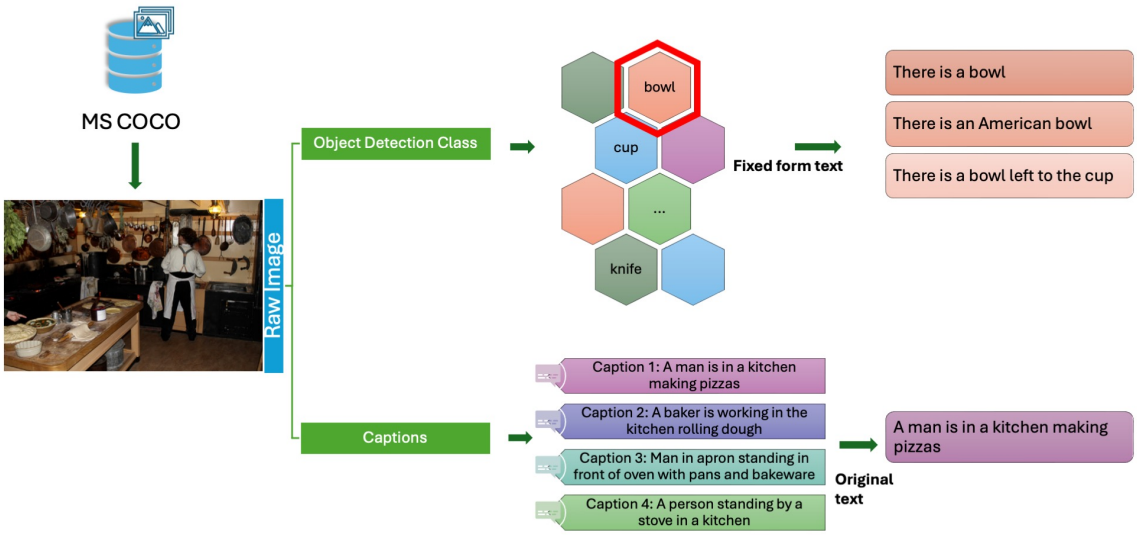

Figure 8: Overview of dataset composition

Table 6: Qualitative summary of CLIPScore and PAC-S behavior across diagnostic axes.

Axis	**CLIPScore / PAC-S Behavior**	**Ideal Metric Behavior**
Visual Properties	Mild penalty on texture/color (CLIPScore more so)	Invariant to superficial visual changes unless semantically meaningful
Text Properties	CLIPScore favors length, complexity / PAC-S favors NEs	Reward informativeness and clarity; avoid verbosity bias
Object Count	Scores slightly decrease with more objects	Fair to complex scenes when captions are accurate
Cultural Context	Default (Culture Neutral) > Cultural modifiers	Culturally neutral scoring for equivalent semantics
Content Category	Domain preference for specific categories like Animal/Appliances over indoor scenes	No unfair preference for content types
Object Size	Scores peak at mid-size (60–80%) objects	Consistent scoring across scales if semantically correct
Spatial Awareness	Slight scoring inconsistency for reversed phrases (CLIPScore)	Equal scoring for equivalent spatial relations
Perturbations	Scores stay high despite incorrect spatial & word order	Strong semantic sensitivity; penalize corrupted captions