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

Anonymous ACL submission

Abstract

Reference-free metrics like CLIPScore and PAC-S are widely used in vision-language tasks, yet their behavior under linguistic, visual, and cultural variation remains poorly understood. We present a systematic audit of these metrics using an eight-factor diagnostic framework applied to 5,000 expert-curated MS-COCO validation images. Across dimensions such as object size, content category, syntax, named entities, spatial relations, and cultural context, we uncover consistent failure modes. Both metrics penalize captions referencing African (5.5%, 4.8%) and Arabian (4.9%, 5.3%) cultures, favor large-object and animal-centric scenes (plus 20 to 30 percent), and show limited sensitivity to spatial negation and word order. These findings reveal cultural and content bias, poor semantic robustness, and weak compositional understanding. We conclude with design recommendations to promote cultural fairness, scale invariance, and semantic grounding in future evaluation metrics for multimodal AI.

1 Introduction

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 & 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. Task integrating vision and language, such as image captioning (Vinyals et al., 2015) and text-to-image (Rombach et al., 2022; Saharia et al., 2022; Ramesh et al., 2021) generation, automatic metrics are often used as standins for human judgment, tasked with assessing the semantic alignment between visual content & text descriptions.

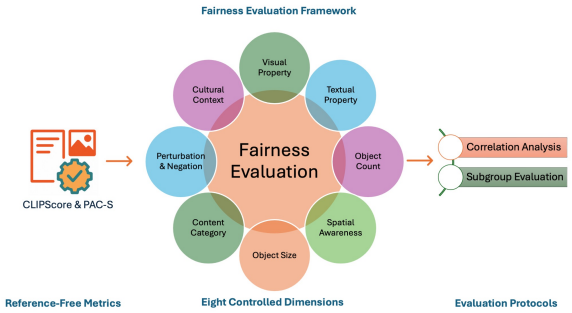

Figure 1: Reference-Free Metrics Evaluation Framework

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 and text evaluation. These metrics compare generated outputs to a fixed set of human-written references and provide interpretable, reproducible scores. However, their reliance on limited reference sets 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 in evaluating flexible or creative generation. In this work, we present a comprehensive audit of **CLIPScore** (Hessel et al., 2021)and **PAC-S** (Sarto et al., 2023a) focusing on their equitability and semantic sensitivity. We use a multidimensional framework combining correlation analysis and subgroup comparisons on the MS-COCO dataset to assess metric responses to scene complexity, syntax, spatial relations, and cultural cues.This reveals consistent biases and misalignments, informing the design of future evaluation tools that are both efficient and fair.

Our main contributions are as follows:

- **Audit of reference-free metrics:** We evaluate popular metrics with a focus on fairness and semantic reliability.

- **Diagnostic framework:** We introduce a multi-dimensional evaluation setup (Fig. 1) spanning linguistic and visual factors.
- **Failure mode insights:** Metrics show bias against African modifiers, human-centric scenes, subtle visuals, and simple syntax.
- **Recommendations:** We propose guidelines to improve future metric development.

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. (2023) 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 Auditing Reference-Free Metrics

To critically assess the reliability and fairness of reference-free evaluation metrics, we propose a systematic evaluation framework that treats the metrics themselves as systems under test. Rather than assuming these metrics to be reliable surrogates for human judgment, we audit them across a diverse set of diagnostic axes designed to reveal hidden bi-

ases, robustness gaps, and semantic insensitivities.

Our analysis focuses on two widely-used reference-free metrics: CLIPScore and PAC-S. We exclude other methods such as UMIC, TIFA, VPEval, and DSG for several reasons. First UMIC metric relies on UNITER a multimodal pretrained model that processes images and text through separate encoding pipelines. This decoupled processing introduces potential alignment issues between modalities and increases complexity in evaluating semantic grounding. Additionally, UMIC's reliance on a pretrained model architecture makes it less transparent and harder to isolate the source of evaluation behavior, which is central to our diagnostic goals. Second, metrics like TIFA, VPEval, and DSG are VQA-based and uses a language models, which introduce additional confounding factors including latency, hardware requirements. More importantly, these methods are ill-suited to our fixed format setup, which centers on single-object scenes. In such cases, only one visually grounded question can typically be formulated, limiting the ability of VQA-based metrics to evaluate fine-grained semantic variation. Consequently, including these methods would distort the scope and validity of our targeted evaluation.

We design our evaluation around a core question: How well do these metrics satisfy key desiderata of a good evaluator? Specifically, we assess:

- **Scene understanding:** Can metrics handle dense, complex, or compositional inputs?
- **Linguistic alignment:** Do metrics reward relevance over verbosity or complexity?
- **Fairness:** Are scores invariant to cultural or contextual variation?
- **Semantic sensitivity:** Do scores reflect correct vs. incorrect text?

3.2 Dataset

We employ the **MS-COCO 2017 validation set** Lin et al. (2015) as our test ground. **MS-COCO (Common Objects in Context)** is an open benchmark extensively utilized in vision-language research with images having bounding box annotations and several human-written captions.

To facilitate controlled evaluation, we used the validation set of MS COCO comprising ∼5,000 images. Images were filtered to contain single dominant objects, identified through bounding box size and category labels. This allowed unambiguous semantic alignment on our diagnostic axes, like spatial relationships, object scale, and cultural modifiers. We constrained the scale intentionally to guarantee:

- Controlled caption-image pairing across several perturbation types, and
- Computational tractability, considering the ∼25,000 instances needed for evaluation.

Captions were drawn from two sources:

- 5 Natural MS-COCO captions, exhibiting natural linguistic variation.
- Fixed-format captions, synthesized through human-authored templates (e.g., "There is a/an [object]", "There is a/an Object A left of Object B" and "There is a/an [cultural_modifier] [object]"). These were used programmatically in order to preserve syntactic coherence & remove semantic change. A

This hybrid setup enabled us to reconcile real-world linguistic variation with controlled experiment, while maintaining reproducibility of the setup and comparability with previous studies based on similarly sized MS-COCO subsets (e.g., González-Chávez et al. (2023), Wu et al. (2018), Kasai et al. (2022)). See Appendix 8 for a visual overview of the dataset construction pipeline.

3.3 Baseline Metric Agreement and Divergence

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. And this suggests that CLIPScore & PAC-S exhibit complementary.

- **T-test:** Reveals statistically significant differences in mean scores.
- **Pearson correlation (r = 0.533):** Indicates moderate alignment, suggesting complementary rather than interchangeable behavior.

Table 1: Statistical comparison between CLIPScore and PAC-S

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

Distributions of metric scores are visualized in Figure 2, showing that CLIPScore is more tightly clustered around central values and scores lower as

compare to PAC-S that exhibits a broader dispersed distribution. This motivates need to evaluate each metric's unique strengths & limitations in detail.

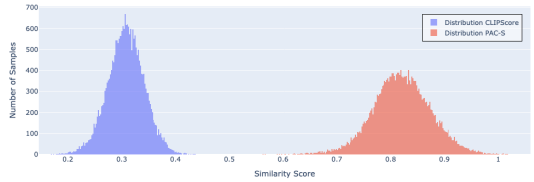

Figure 2: Score distribution comparison between CLIPScore and PAC-S.

3.4 Evaluation Protocol

We use two strategies to analyze metric behavior:

Correlation Analysis: Spearman & Pearson correlations are computed between metric scores & visual/textual properties (e.g., object count, color variance, caption complexity). Strong correlations in unintended directions reveal potential biases.

Subgroup Comparison: We compare average scores across controlled subgroups (e.g., "American" vs. "African", or "small" vs. "large") to assess fairness and semantic consistency.

Both metrics are evaluated across eight diagnostic dimensions (Table 2).

4 Results

4.1 Textual Property - Evaluating Sensitivity to Linguistic Structure

An ideal evaluation metric should focus on how well a caption matches the meaning of the image, rather than on things like sentence length or how complex the wording is. In the section shown in Figure 3, we check whether CLIPScore and PAC-S rely too heavily on surface-level language features. To do this, we measure how strongly each of following attributes is related to scores assigned by each metrics:

1. **Text Length**: Measured as total number of non stopword tokens
2. **Sentence Complexity**: Defined as a ratio of tokens and noun phrases to the number of clauses, approximating syntactic density.
3. **Flesch–Kincaid Grade Level**: Indicating U.S. school grade level required to understand a text.
4. **Named Entity Count**: Captures the number of named entities (e.g., people, places, organizations) identified in the caption.

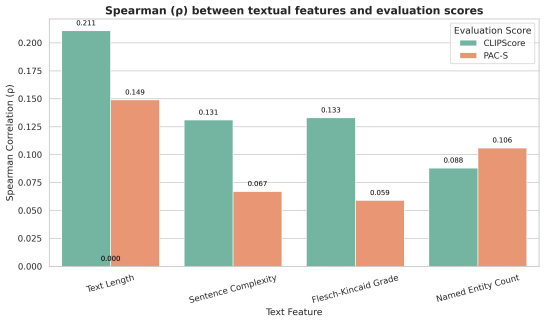

Figure 3: Spearman (ρ) between textual features and evaluation scores.

Observation: CLIPScore exhibits a notable bias toward longer captions ($r = 0.211$), suggesting a preference for verbosity over semantic precision; PAC-S also shows a positive but weaker correlation ($r = 0.149$). Additionally, CLIPScore correlates positively with syntactic complexity and readability grade, indicating sensitivity to caption structure, while PAC-S appears less affected. Interestingly, PAC-S shows slightly greater responsiveness to the presence of named entities ($r = 0.106$), reflecting a modest entity preference.

Metric Expectation: Metrics should score captions based on semantic relevance.

Failure Mode: CLIPScore favors complex sentences, often penalizing concise but accurate ones, while PAC-S prefers simpler language but is biased toward named entities.

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

A reliable evaluation metric should remain invariant to superficial visual variations that do not alter semantic meaning. In Figure 4, we assess whether CLIPScore and PAC-S are unintentionally influenced by low-level visual properties of images. To do this, we analyze how each of the following image attributes correlates with the scores they assign:

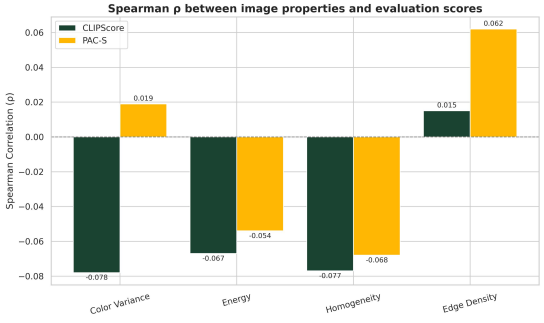

Figure 4: Spearman ρ between image properties and evaluation scores.

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

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.

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 affecting image semantics.
3. **Edge Density:** Calculated using the Canny edge detector as the ratio of edge pixels to total image pixels, indicating visual detail.

By correlating these measures with CLIPScore and PAC-S, we assess whether the metrics remain robust to superficial visual variability.

Observation: 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 uniform images irrespective of semantic accuracy. Additionally, PAC-S shows a slight preference for images with higher edge density ($r = 0.062$), reflecting a bias toward more detailed or structured visuals, whereas CLIPScore remains mostly invariant.

Metric Expectation: Scores should remain stable across variations in color, texture, or edge structure unless they impact the caption's correctness.

Failure Mode: Both CLIPScore and PAC-S exhibit minor but systematic visual sensitivity, suggesting that they partially conflate stylistic or perceptual features with semantic quality.

4.3 Object Count – Assessing Compositional Generalization

A reliable evaluation metric should handle complex scenes with multiple objects, as commonly found in real-world settings like surveillance or robotics. Captions for such images should not be penalized due to scene complexity.

Table 3 examines whether CLIPScore and PAC-S are sensitive to object count by correlating their scores with the number of distinct object classes per image, computed from MS-COCO annotations.

Table 3: Spearman (ρ) between object count and evaluation scores.

Feature	CLIPScore	PAC-S
Object Count	-0.084	-0.080

Observation-Negative Correlation with Complexity: Both metrics exhibit a small but consistent negative correlation with object count, suggesting that as the number of objects in a scene increases, evaluation scores tend to decrease.

Metric Expectation: Metrics should treat correctly grounded, multi-entity captions fairly across all salient objects.

Failure Mode: Both CLIPScore and PAC-S penalize complex images with multiple entities, revealing a limited capacity.

4.4 Cultural Context – Auditing Cultural Fairness in Evaluation

A good evaluation score should be culturally agnostic-giving similar scores to semantically identical captions, irrespective of geographical or cultural modifiers. To check for this, we experimented with how CLIPScore and PAC-S react to captions which only vary by cultural adjectives, while the object identity and syntax are kept constant.

For every image with a single object (e.g., "chair", "car"), we created fixed-syntax captions like: *"There is a/an [American/African/Asian/European/Russian/Arabian*

/Oceania] [Object_name]" We present the average scores given by each metric in Figure. 5

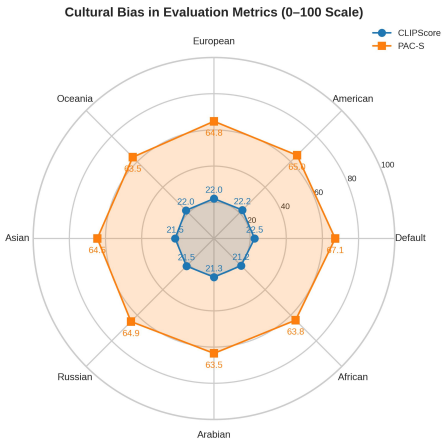

Figure 5: Cultural bias analysis using a radar plot showing evaluation scores (on a 0-100 scale) across various cultural regions.

Observation: Both CLIPScore and PAC-S consistently assign lower scores to culturally modified captions compared to the default, indicating a uniform drop in performance across modifiers. CLIPScore shows the strongest bias against African (-5.5%) and Arabian (-4.9%) descriptors, while PAC-S registers similarly steep declines for Arabian (-5.3%), Oceania (-5.2%), and African (-4.8%) modifiers. In contrast, American and European references receive scores closest to the baseline, revealing a clear Western bias present in both evaluation measures.

Metric Expectation: Scores should reflect only the correctness and grounding of visual-text alignment, not the cultural identity or geographic descriptor of an object.

Failure Mode: CLIPScore and PAC-S both demonstrate systematic Western cultural descriptor bias, even with identical syntactic templates.

4.5 Object Category – Evaluating Content-Type Sensitivity

A fair evaluation metric must look at semantic correctness irrespective of the nature of the content illustrated whether it's an animal, a human, an object, or an element of the scene. Systematic scoring bias in favor of some content types, without semantic grounds, reflects domain-level bias.

To evaluate this, we considered average metric scores over 12 MS-COCO supercategories, employing fixed-formatted captions to images with a single dominant object from each supercategory. CLIPScore mean and PAC-S mean are depicted in Figure. 6.

Observation: Animal-related content consistently receives the highest scores from both metrics (CLIPScore: 0.2508; PAC-S: 0.7341), aligning with prior reports of animal bias. Appliance and sports categories also score relatively high, while person, kitchen, and accessory categories receive the lowest scores. Notably, person-class objects show the greatest negative deviation, with CLIPScore 16.2% and PAC-S 11.6% below the mean, highlighting a substantial undervaluation in both metrics.

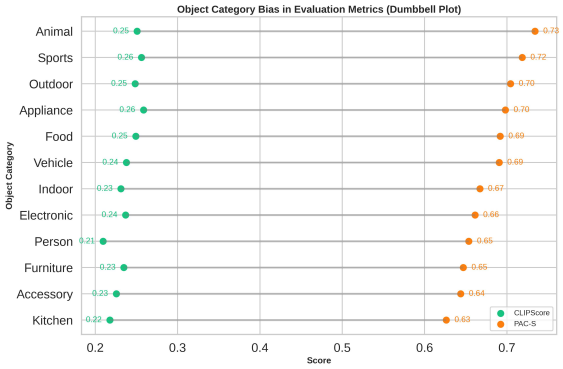

Figure 6: Dumbbell plot showing Object Category bias, indicating metric sensitivity to semantic content.

Metric Expectation: Metrics should evaluate captions consistently across content types when the semantic match is equivalent.

Failure Mode These findings indicate content-type bias, with person-centric and indoor scenes undervalued, while animals, appliances, and sports items are consistently over-scored likely reflecting pretraining data biases.

4.6 Object Size – Evaluating Scale Sensitivity and Visual Prominence Bias

An effective evaluation metric should be scale-invariant assigning similar scores to correct captions regardless of object size. Otherwise, it may undervalue small-object recognition or penalize captions in cluttered or zoomed-out scenes.

In Figure 7 We group captions by the object's image area percentage & compute average scores using fixed-form captions, isolating impact of object size while keeping caption structure constant.

Observation: Evaluation scores increase with object size, showing a clear sensitivity to scale. Both

metrics peak in the 60–80% size 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 difficulties in grounding captions to fine details, while extremely large objects (90–100%) also perform poorly, possibly due to loss of contextual grounding in overly cropped or zoomed-in images.

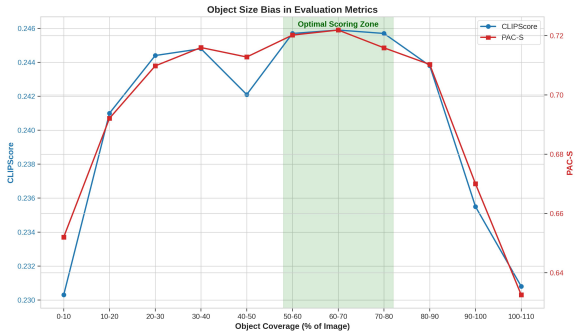

Figure 7: Evaluation metrics (CLIP and PAC-S) peak within the 50-80% object coverage range, indicating a bias toward medium-sized objects.

Metric Expectation: Correct captions should receive consistent scores on object scales, semantic correctness should not be penalized by prominence.

Failure Mode: Both metrics favor mid-sized objects and undervalue captions about very small or large ones, limiting their reliability in tasks requiring fine-grained visual grounding.

4.7 Spatial Awareness – Testing Positional Sensitivity and Object Relations

Evaluation metrics should be invariant to absolute object placement and treat equivalent spatial relationships equally. For instance, "object1 is to the left of object2" and "object2 is to the right of object1" should receive similar scores.

Table 4 assesses whether CLIPScore and PAC-S meet this standard, examining their sensitivity to both absolute and relative spatial positioning.

We use a fixed format based captioning approach to systematically isolate spatial variables, as described below:

1. **Absolute Positioning:** To evaluate positional bias, we compare scores for identical captions describing objects on different sides of the image. For left-side scores, we use the original image (if the object is on the left) or horizontally flip it (if the object is on the right). The process is reversed to compute right-side scores, ensuring that only object position changes while caption remains fixed.

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%)**

2. **Relative Positioning:** We create pairs of captions describing the same spatial relation in different orders. For an object pair appearing in the sequence `(object_i, object_j)`, we generate the following captions:
   - There is a/an [object_i] left to [object_j]
   - There is a/an [object_j] right to [object_i]

Observation: Both metrics demonstrate robustness to absolute positioning, showing nearly identical scores for objects placed on the left or right side of the image. However, in relative positioning scenarios, CLIPScore shows a slight preference for "right of" relations, revealing minor inconsistencies in handling directional spatial descriptions.

Metric Expectation: Evaluation metrics should be invariant to absolute positioning, assigning similar scores whether an object appears on the left or right. They should also treat equivalent relative expressions (e.g., "A is left of B" vs. "B is right of A") as semantically identical.

Failure Mode: Minor asymmetries in CLIPScore assignment for relative spatial descriptions suggest potential model biases or sensitivity to language formulation.

4.8 Perturbations & Negations

Robust evaluation metrics should distinguish between captions that are spatially and semantically correct and those that contain errors. In Table 5, we test whether CLIPScore and PAC-S can: penalize spatially incorrect captions, remain unaffected by irrelevant visual changes, and detect syntactic degradation in text. We evaluate the metrics using the following types of perturbations and negations:

Spatial Negation Sensitivity: We evaluate metric sensitivity to spatial errors using two types:

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

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 %**)

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).
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).

Multimodal Input Perturbations: We apply the following transformations to assess metric robustness: Convert images to grayscale, Shuffle caption word order and Reverse caption word order.

Observation: CLIPScore often fails to penalize spatially incorrect captions, sometimes scoring them higher than correct ones. PAC-S performs slightly better but with minimal margin. Both metrics show resilience to visual changes like grayscale conversion, and limited sensitivity to syntactic disruptions maintaining relatively high scores even with shuffled or reversed captions, indicating a bag-of-words behavior.

Metric Expectation: Penalize semantically incorrect captions. Maintain robustness to irrelevant visual changes. Reflect decreased alignment when sentence structure is syntactically degraded.

Failure Mode: CLIPScore and PAC-S show low sensitivity to semantic corruption, relying more on keyword overlap than true meaning, even with negated or disordered captions.

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

In sum, existing reference-free measures are lacking in fairness, semantic sensitivity, and robustness preventing their use for assessing varied, real-world captioning outputs.

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 work offers a thorough review of reference-free measures, it is limited by the following methodological decisions. We used single-

object images to facilitate controlled experimentation, precluding direct applicability to real-world, multi-object scenes. The dataset size ($\approx$ 5,000 images) was also kept small for computational tractability ($\approx$ 25,000 evaluations), consistent with previous works but restricting generalizability to larger or more heterogeneous datasets. Moreover, we only tested two metrics CLIPScore and PAC-S leaving other new methods like VQA-based or LLM-based scoring out of consideration because of compatibility limitations.

Our cultural fairness audit, although more encompassing than in prior work, was restricted to seven modifiers and only seven global regions (e.g., Latin America, Indigenous populations). Further, the employment of fixed-format captions, which is convenient for discounting semantic change, does not capture the richness of naturally occurring writing. These approximations can affect how metrics handle more representative variation in language. Future research should remedy these limitations to enable more thorough, inclusive, and ecologically valid assessments.

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.

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.

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.

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*.

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*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *arXiv preprint arXiv:2102.12092*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

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.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Ruslan Salakhutdinov, David Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

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.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 3156–3164.

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.

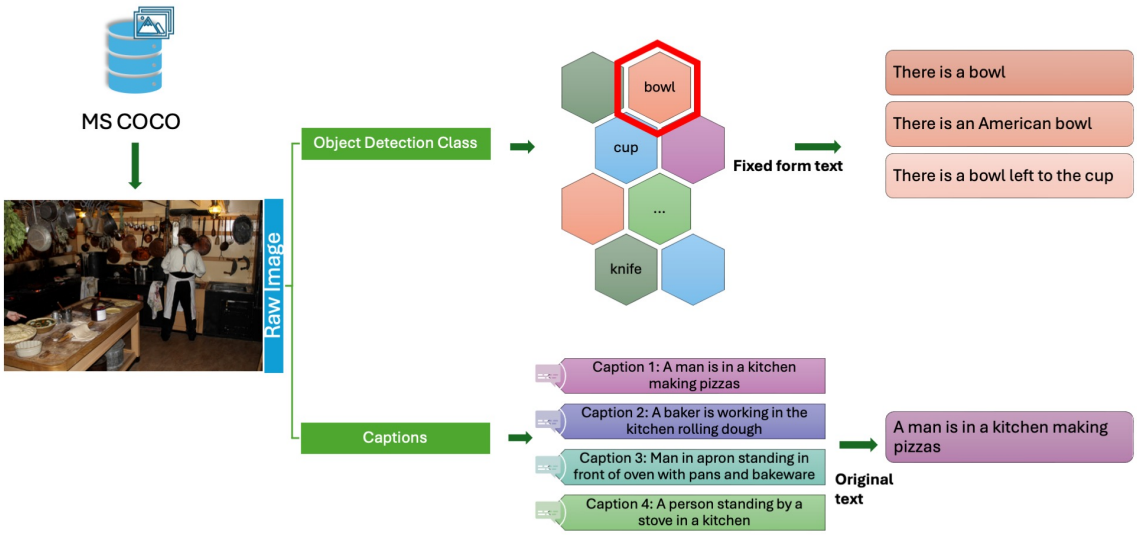

Figure 8: Overview of dataset composition

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