Understanding Human Heuristics in Context-Sensitive Image Captioning

Yanru Jiang
yanrujiang@g.ucla.edu
University of California, Los Angeles
Department of Communication
Department of Statistics
USA

Hongjing Lu
hongjing@g.ucla.edu
University of California, Los Angeles
Department of Psychology
Department of Statistics
USA

Rick Dale
rdale@ucla.edu
University of California, Los Angeles
Department of Communication
USA

Abstract

Recent studies highlight the context sensitivity of image captioning, where the context in which an image appears strongly influences its caption's informativeness and linguistic style. While AI-generated text increasingly mirrors human language, its informativeness, derived from cross-modal image-text reasoning, may still fall short of expert-authored content. Given the intertwined nature of informativeness and linguistic style, this study examines news image captioning, a naturally high-context task, to manipulate caption informativeness and assess human sensitivity to such variations. Two experiments (N = 378) and a series of logistic regression analyses reveal that while humans effectively interpret informational cues, their intuition about AI linguistic style often diverges from actual AI markers. Moreover, humans more readily integrate multiple modalities in preference tasks but rely heavily on linguistic-based strategies for AI detection. These findings underscore the adaptability of human evaluation in image-text systems and suggest informative signals as the more reliable basis for judgment.

CCS Concepts

• **Human-centered computing → Empirical studies in HCI**.

Keywords

Generative AI, Multimodal Communication, Language-Vision Models, AI-Mediated Communication (AI-MC), Image Captioning

ACM Reference Format:
Yanru Jiang, Hongjing Lu, and Rick Dale. 2025. Understanding Human Heuristics in Context-Sensitive Image Captioning. In *Proceedings of the Human-Centered Evaluation and Auditing of Language Models (HEAL) Workshop at CHI Conference on Human Factors in Computing Systems (CHI '25).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Before the advent of large vision-language models (LVLMs), image captioning was a nontrivial task for machines, as it involved sophisticated visual recognition, language generation, and cross-modal reasoning [1]. Beyond these technical challenges, recent studies have emphasized that image captioning is highly context-sensitive—where the context in which an image appears can significantly influence both the expected informativeness and linguistic style of a caption [14].

While a comprehensive evaluation of image captions could consider a range of linguistic, visual, and cross-modal features, this study specifically focuses on the linguistic and informational dimensions. The emphasis on linguistic style is motivated by prior findings that modern AI systems can closely mimic the linguistic patterns of human-generated text [5, 18], while human heuristics for detecting such stylistic differences are often unreliable [5]. These dynamics, previously explored in text-only AI-generated content, are equally relevant in the image-to-text setting and warrant further investigation. In contrast to linguistic style, informativeness is a feature more uniquely tied to image captions—especially in the context of news media, where audiences must rely on captions to extract information beyond the image itself. This may include alignment with the visual content, retrieval of contextual details from the accompanying article, or identification of named entities such as people, places, or events depicted in the image [16].

Given the intertwined nature of informativeness and linguistic style in image captioning, this study examines news image captioning—a naturally high-context task—to manipulate caption informativeness and assess human sensitivity to such variations. We operationalize informativeness and linguistic style by extracting measurable features informed by prior research [5, 12, 16] and validate this operationalization using stepwise logistic regression. Next, we assess the reliability of human judgments in evaluating these two dimensions. Finally, we examine how human sensitivity to linguistic style and informativeness shifts between evaluation and AI detection tasks.

1.1 Image Captioning as a Generative Task

Automatic image description is a highly challenging task that requires machines to perceive and recognize various visual elements (such as objects, actions, attributes, and scenes), understand their compositions and semantic relationships, and generate linguistically coherent descriptions that align with human cognition [1, 15]. This complexity and the interdependence between visual and textual semantics have attracted interest from both the computer vision (CV) and natural language generation (NLG) communities in the past. From a CV perspective, image descriptions can refer to multiple levels of visual elements within an image, ranging from

CHI '25, Yokohama, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

a person's identity to the clothing they wear. From an NLG standpoint, models must translate these detected visual components (i.e., non-linguistic representations) into human-like linguistic expressions, adhering to cultural conventions regarding description length, word choice, sentence structure, and information prioritization. Unlike simple image descriptions that focus on verbalizing what is visually present, image captioning poses a greater challenge due to its large solution space, requiring models to engage in visual storytelling and convey contextual information beyond what is explicitly visible in the image [1].

Previous research highlights that automatic image description and captioning is not a one-size-fits-all task [11]; the information needs and linguistic style of captions shift depending on the context in which an image appears. The same image may be described differently across domains such as news, social media, e-commerce, employment websites, or academic publications [14]. For example, an image of a celebrity might prompt a focus on their attire in a fashion editorial, whereas a news article would prioritize their identity and relevance to an event [10]. Similarly, stylistic expectations vary—social media captions often adopt a personal tone, while news captions trend to follow journalistic conventions. These variations highlight the difference between assessing captions for informativeness and for linguistic style.

1.2 Evaluating Image Captions

Recent advancements in LVLMs have demonstrated impressive capabilities in both visual reasoning and language generation. These models can easily adapt to different linguistic styles through prompting, mimicking journalistic, conversational, or descriptive tones with minimal effort [5, 12]. However, while they excel at replicating linguistic conventions, their ability to perform complex image-text reasoning beyond stylistic adaptation in image captioning tasks remains underexplored.

A common computational approach for evaluating automatic image description and captions is the use of referenceless metrics, such as CLIPScore [4], which assesses image-caption similarity using pre-trained vision-language models [13] without requiring ground-truth labels. While these metrics offer efficiency and scalability, they do not explicitly account for informational appropriateness, linguistic preference, and context-sensitivity [6].

The challenge in evaluating the linguistic style and informativeness of AI-generated captions is that these factors are deeply intertwined. For instance, the presence of proper nouns (e.g., names of locations or public figures) can signal higher informativeness by providing specific contextual references [19], yet it may also reflect stylistic tendencies favoring more descriptive language. This entanglement complicates efforts to measure the influence of contextual information on AI-generated captions and human judgments of their quality. To address these gaps, the current study employs an experimental design that that manipulates the informativeness of captions based on the presence of image context (image-only vs. image + article), and investigate if humans are sensitive to these different factors.

1.3 Context, Informativeness, and Linguistic Style in Evaluating News Image Captions

News media today distribute information globally through various modalities, including text, images, audio, and video [2]. Recent advances in generative AI [3, 4] have enabled the creation of AI-generated captions that closely resemble journalist-written ones [8], with minimal technical barriers.

News image captioning is a naturalistic cross-modal reasoning task that places high demands on a model's world knowledge, requiring it to recognize or infer information about people, locations, and events beyond visually grounded entities [12]. Compared to other naturalistic captioning tasks, such as social media posts, news captions are typically based on recognizable figures or events and follow structured linguistic norms shaped by journalistic conventions, such as who, when, where, and what (misc) [16]. Thus, access to high-quality contextual information serves as a key factor in the informativeness of AI-generated captions. Previous NLG research has shown that providing models with article content enhances caption quality in a trackable way by supplying both visually grounded entities for "who" and "where" and non-visually grounded information like "when" and "misc" [9, 16]

Therefore, this study selects news image captioning to compare AI-generated captions under two conditions: image-only vs. image + article. Without article access, LVLMs rely on internal knowledge to supplement missing details, whereas with article access, AI-generated captions are expected to improve by incorporating named entities and event-specific information.

To separate informativeness from linguistic style, we define informativeness as the effective integration of an image and its caption. We measure this using CLIP image-caption similarity and the presence of named entities—specifically mentions of "who" and "where," following journalistic conventions [12, 16]. Linguistic style, on the other hand, is defined by a set of AI language markers identified in prior AI detection studies [5]. Next, we computationally extract both informational and linguistic features to analyze their role in distinguishing between journalist-written and AI-generated captions. In this step, we conduct a confirmatory analysis, expecting linguistic features to be strongly associated with all AI-generated captions but not to reliably differentiate between AI captions generated with or without article context. On the other hand, informational features should strongly correlate with AI captions generated using article content. Once we validate these feature classifications, we examine whether humans can reliably use these cues when evaluating news image captions in two tasks: caption preference and AI caption detection. We also analyze how their reliance on these features shifts between the two tasks.

2 AI-generated News Image Captions

2.1 Sampling Image, Caption, Article Pairs

Image-caption stimulus pairs for all experiments were generated from Voice of America (VOA) news, one of the oldest and largest U.S.–funded international broadcasters, collected by Li et al. [7]. This VOA dataset contains 1,014 new image-caption pairs along with 199 accompanying articles. Each article comes with one to eight images with original captions created by VOA journalists. To

acquire the ⟨*image, caption, article*⟩ stimulus set for *with-article* vs. *without-article* conditions, one image-caption pair was randomly selected from each of the 199 articles. A total of 136 sets were selected for AI caption generation, ensuring a balanced representation across news topics, while excluding extreme images (e.g., those depicting violence or dead bodies) to comply with IRB requirements.

2.2 Generating AI Captions

The AI-generated captions for *without-article* conditions are generated using GPT-4Vision, given the corresponding image. The AI-generated captions for *with-article* conditions are generated by the same models given both the corresponding image and news articles provided by the VOA dataset. An example prompt for AI models is: *"Generate a VOA news caption for the given image, [based on the following news article: '{article}'] in the style of Voice of America (VOA) news reports. Keep it around 25 words."* The count of 25 words was calculated based on the average number of words in the human captions. Without such guidance, the model might generate captions with several sentences, diminishing the ecological validity of the comparison. The AI captions were generated over multiple days in November 2023 and February 2024 due to usage limits. There is no temperature hyperparameter in the GPT-4Vision model.

3 Human Evaluation of Image Captions

Using a news image caption dataset, we conducted a series of experiments and analyses to examine human preferences for AI-generated versus journalist-generated captions (Experiment 1), their ability to correctly identify AI-generated captions (Experiment 2), and the factors influencing their preferences across various contexts, including differing article presence, through the computational extraction of linguistic and cross-modal features. The journalist-generated captions are the original captions provided by the VOA dataset. All experiments were approved by the Institutional Review Board (IRB Protocol #24-000019, February 2024).

3.1 Participant Recruitment

The collected data include participants' choices among an original (journalist-generated) caption and an AI-generated caption for each image-caption pair as well as their demographic information. Experiment 1 (N = 192) and Experiment 2 (N = 186) were collected from participants recruited through the subject pools. All participants were undergraduates. There was no overlap between participants across experiments, subsets of stimuli, or between the article and no-article conditions. Undergraduates were an appropriate demographic for our study, as they are generally familiar with digital media and AI-generated content, making them reasonably equipped to assess these captioning tasks.

3.2 Procedure

Each experiment uses a between-subject design to examine how participants choose captions based on images, with or without contextual information. In each condition, participants conduct two-alternative forced choice (2AFC) tasks for a total of 136 stimulus sets, where for each stimulus, they see two captions, one from a journalist and one from AI, and select the better one. The journalist's original caption serves as a baseline for evaluating how humans assess AI-generated captions.

In all experiments, participants were first presented with a set of instructions and two exemplar questions to familiarize them with the task. Three attention checkers, each including one caption that is obviously not relevant to the corresponding image and one original human-generated caption, were added to the survey and randomized with other image-caption pairs. Participants who could not pass two out of three attention checks were excluded from the analysis.

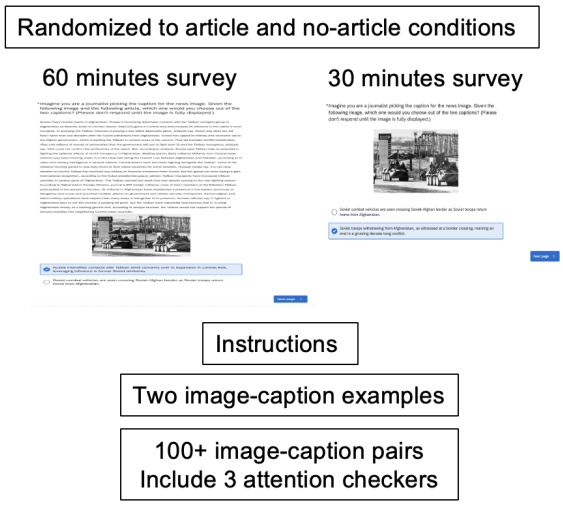

Figure 1: An illustration of between-subject conditions for Article and No Article conditions. In the Article condition, the article context is provided to both participants and the model for caption generation. Images adopted from [7].

3.3 Experiment 1 Human Preference

Experiment 1 compared human preferences for choosing AI-generated captions when articles are provided (to both the audience and AI models) and when they are not. This experiment asked "Imagine you are a journalist picking the caption for the news image. Given the following image (and the following article), which one would you choose out of the two options?" in a 2AFC setting for each image-caption pair. We recruited 200 participants across the two conditions, with 192 passing attention checks and included in Experiment 1.

3.4 Experiment 2 Detecting AI Captions

Experiment 2 aims to assess whether the audience can distinguish AI-generated captions from the human-generated ones, both with and without the article being provided (same settings as Experiment 1). Specifically, it asks "Considering the given image (and the article), one choice is created by a human, and the other by AI. Which option do you think was generated by AI?" We recruited a total of 200 participants across both conditions, with 186 passing attention checks and included in Experiment 2.

3.5 Computational Identification of Informational and Linguistic Features

To understand the underlying decision-making process when audiences encounter image captions in different contexts, we computationally extracted a wide range of linguistic and cross-modal features from both journalist- and AI-generated captions (with and without articles), following a similar approach to Jakesch et al. [5].

Linguistic features were defined based on those identified by Jakesch et al. [5] through machine learning-based feature selection and are relevant to caption generation tasks. The initial list of features includes *Word Count*, LIWC dictionaries (such as *Affect, Past Tense, Pronouns, Conjunctions, Causation, Differentiation, Quantifiers, Adverbs*, etc.), and *Proper Noun*.

Regarding informational features, the integration of visual and textual semantic coherence is assessed using *CLIPScore*, which measures the semantic similarity between an image and its caption. Since news image captioning tends to follow the journalistic convention of ⟨who, when, where, what⟩ [16], we additionally utilized SpaCy for named entity recognition (NER) to detect *WHO* (including PERSON, NORP, ORG) and *WHERE* (including FAC, GPE, LOC). These named entities have been observed to improve caption quality when articles are provided, making them suitable features for examining informational enrichment from article context [9, 19]. WHO and WHERE are also understood as more visually grounded entity types compared to WHEN and WHAT, making them strong indicators of cross-modal integration. Considering the large number of variables, we employed stepwise feature selection in the following analysis to eliminate variables with minimal contributions to model performance.

4 Results

4.1 Human vs. Model Judgment

To compare human and model assessments at an aggregated level, we report subject-level caption preference and human-likeness, alongside two model-based measures—semantic alignment between the image and caption (measured by CLIPScore) and information retention in the caption given the article context (measured by BERT Recall Score)—which serve as naïve judgments based on single-dimensional evaluations (see Figure 2).

4.1.1 Human Judgment. Human assessments were computed by averaging the proportion of participants who chose AI-generated captions over the original ones across 136 stimuli, based on responses from Experiment 1 (preference) and Experiment 2 (perceived as AI). To align the directional impact of article access across judgments, human-likeness was calculated by reverse coding the "detected as AI" responses in Experiment 2.

An independent t-test revealed that the audience significantly prefers AI-generated captions over human-generated ones when both the audience and AI have article context compared to when no context is provided (t = 3.16, p = 0.002). Under the no-article condition, audience preference is at chance level (μ = 53.49%, SD = 1.47%), indicating that the audience does not differentiate between the two types of captions when only the image is provided and context is lacking. Under the article condition, the audience systematically prefers the AI caption over the original caption (μ = 61.33%, SD = 2.02%). These results suggest that current SOTA models can achieve at least a reasonable level of integrative image-text reasoning, similar to human generators, in their linguistic generation capabilities.

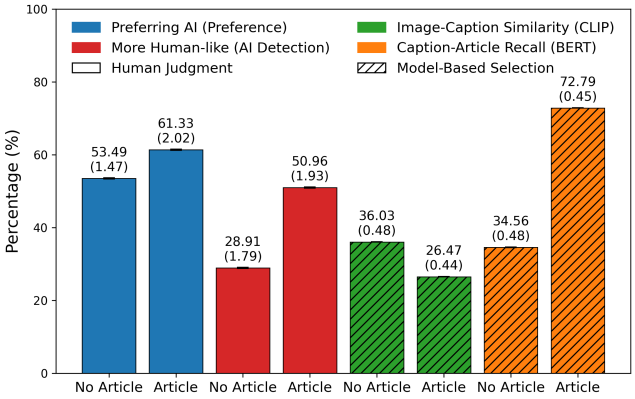

Figure 2: Comparison of Human and Model-Based Judgments Across Conditions. The figure illustrates the percentage of selections for Preferring AI-Generated Captions, More Human-like (1 - Detected as AI%), Image-Caption Similarity (CLIP-Score), and Caption-Article Recall (BERT Recall Score between caption and article) across No Article and Article conditions.

Experiment 2 confirms that participants cannot distinguish AI-generated captions from human captions when context is provided to AI (μ = 50.96%, SD = 1.93%). When no article is provided to either the model or the audience, participants can identify AI-generated content (i.e, they perceived the original caption as more human-like 28.91% of the time; SD = 1.79%).

4.1.2 Model-Based Selection. Examining model-based assessments, we observe the limitations of computational metrics, as discussed previously. Looking at CLIPScore-based judgments, we find that under both the no-article condition (μ = 36.03%, SD = 0.48%) and the article condition ((μ = 26.47%, SD = 0.44%), the results do not correspond to human judgments. This suggests that if we only consider image-text semantic alignment, we might erroneously conclude that generative models struggle with visual recognition and cross-modal linkage. However, human evaluations involve multidimensional assessments of image-caption fit, as image captioning is designed to provide visual storytelling and introduce external information to aid comprehension. Even more concerning, providing an article, which helps generative models incorporate external information beyond the image (a desirable feature in many image-captioning tasks), can actually lower CLIPScore. This further reveals its deviation as a reliable image captioning metric.

While CLIPScore's limitation lies in only considering image-text similarity, BERT Recall Score [17] better captures the directional impact of article context, as it directly measures information retention in captions. However, it remains a coarse measure compared to human assessments, since access to an article may introduce

noise in captions or mislead AI generators regarding the relevance of information to the image.

These distinctions between human and model judgments, along with differences in how each measure responds to article context, highlight the complexity of news image captioning tasks. These findings motivate further investigations into decomposing the multidimensional nature of human evaluation, particularly by examining caption informativeness and linguistic style.

4.2 Informational and Linguistic Features in Captions

We first conducted logistic regression with stepwise feature selection to identify features significantly associated with captions being AI-generated, compared to journalist-written captions (Model 1 in Table 1). Across all captions (both human-written and AI-generated, with or without article context), both informational named entity features—WHO entities (OR = 0.48, 95% CI: [0.34, 0.67]; $p < .001$), and WHERE entities (OR = 0.69, 95% CI: [0.52, 0.92]; $p < .05$)—as well as certain linguistic markers, significantly associated with captions being AI-generated. These linguistic markers, termed *AI Linguistic Markers*, included Affect Words (OR = 2.55, 95% CI: [1.83, 3.68]; $p < .001$), Past-Focused Words (OR = 0.46, 95% CI: [0.33, 0.62]; $p < .001$), Pronouns (OR = 0.60, 95% CI: [0.45, 0.78]; $p < .001$), and Conjunctions (OR = 1.97, 95% CI: [1.44, 2.76]; $p < .001$).

Next, we conducted a similar regression focusing exclusively on AI-generated captions to identify features distinguishing captions created with or without article context (Model 2). Informational features significantly associated with captions generated with article access—CLIP Similarity (OR = 1.67, 95% CI: [1.23, 2.32]; $p < .001$), WHO entities (OR = 0.24, 95% CI: [0.15, 0.36]; $p < .001$), and WHERE entities (OR = 0.33, 95% CI: [0.22, 0.48]; $p < .001$). In contrast, linguistic markers were not significantly associated with article access (Affect Words: OR = 0.79, 95% CI: [0.57, 1.09], Past-Focused Words: OR = 1.15, 95% CI: [0.79, 1.68], Pronouns: OR = 0.85, 95% CI: [0.62, 1.15], Conjunctions: OR = 1.19, 95% CI: [0.88, 1.61]). This confirms that informational features, as computationally identified, are sensitive to the additional context provided by the article, while linguistic features remain unaffected by article access in AI-generated captions.

Comparing Models 1 and 2, these findings support the assumption that article access improves caption informativeness, bringing AI-generated captions closer to journalist-generated captions in a measurable way. Specifically, article access improves named entity coverage and localization for figures and locations. One might suspect that proper nouns could also serve as indicators of informativeness, given their strong correlation with named entities. Indeed, our feature selection models support this intuition: when both proper nouns and named entities were included, they exhibited strong collinearity. However, named entities were retained more frequently than proper nouns in the feature selection process, suggesting their stronger contribution.

On the other hand, features that are nonsignificant in Model 2 but significant in Model 1 help validate the identified context-independent AI linguistic markers. These features do not differentiate between AI-generated captions with or without access to additional context but reliably distinguish AI-generated captions from human-written ones. Such markers represent inherent characteristics of AI-generated text that remain consistent, even when the input information provided to the AI for caption generation varies greatly.

4.3 Informational and Linguistic Cues in Human Evaluation

Once we categorized features into informational features and AI linguistic markers based on their association with different AI-generated captions, we conducted a logistic regression with mixed effects and interaction effects (article presence × informational features), combined with stepwise feature selection. This analysis aimed to identify which features predict AI detection (Model 4) and human preference (Model 3) of image captions. Individual random effects were included in the mixed-effect models to account for participant-specific variations. To ensure alignment in the direction of human assessment (e.g., captions perceived as AI-generated aligning with lower human preference), we reversed the "caption preference" scale to "less preferred" captions.

Comparing Model 1 (actual AI-generated captions) with Model 3 (captions perceived as AI-generated), we observed that human participants' reliance on informational cues was consistent with the informational features associated with AI-generated captions, as these features exhibited the same directional associations in both models. This alignment was stronger for person entities (WHO: OR = 0.88, 95% CI: [0.84, 0.92]; $p < .001$) than for location entities (WHERE: OR = 0.97, 95% CI: [0.94, 1.01]; $p < .05$), suggesting that human perceptions of news image captioning tend to emphasize figures more than locations.

However, human assessments of linguistic cues showed weaker or reversed directional associations compared to the computationally identified AI linguistic markers in Model 1. Participants also relied on additional linguistic heuristics that were not significantly associated with AI-generated captions in Model 1. We categorize these as *Human Heuristic* features, reflecting patterns that humans intuitively associate with AI but that do not consistently differentiate AI-generated text.

A similar pattern emerged when comparing Model 1 (actual AI-generated captions) with Model 4 (caption preference): both cross-modal and person-entity informational cues showed consistent directional associations between Model 1 and Model 4 (OR = 0.79, 95% CI: [0.77, 0.82], $p < .001$; OR = 0.93, 95% CI: [0.89, 0.97], $p < .001$). However, the directional association was opposite for location entities (OR = 1.07, 95% CI: [1.04, 1.10], $p < .001$). These similar patterns indicate that the features humans associate with captions being more human-like (i.e., less AI-generated) are generally linked to how they evaluate caption quality. However, these associations often diverged from the actual characteristics of AI-generated captions, highlighting discrepancies between human intuition and the actual linguistic markers of AI-generated text.

These findings suggest that while humans reliably associate the correct signals indicating that caption generators are more informative by incorporating contextual information, they struggle to accurately associate the linguistic styles of current AI systems. The discrepancy between informational and linguistic cues may partially explain why humans tended to prefer AI-generated captions

Table 1: Odds Ratios for Context-Related and Context-Independent Features. Significance levels: * < 0.05, ** < 0.01, * < 0.001. Models (1) and (2) include all captions, with Model (1) covering all captions and Model (2) focusing solely on AI-generated captions. Models (3) and (4) analyze human assessments from Experiments 1 and 2, with Experiment 1 responses reverse-coded to align with the interpretation of other regressions. All regression models used feature selection, excluding minimally contributive features with negligible impact on coefficient estimation. Article interaction effects were applied only to informational features, as informativeness was manipulated based on article presence.**

Model	Actually AI-generated (1)	Without Article Access (2)	Perceived as AI (3)	Less Preferred (4)
Informational Features				
CLIP Similarity	0.77 (0.57, 1.03)	1.67 (1.23, 2.32)**	0.96 (0.93, 0.99)**	0.79 (0.77, 0.82)***
Named Entities (WHO)	0.48 (0.34, 0.67)***	0.24 (0.15, 0.36)***	0.88 (0.84, 0.92)***	0.93 (0.89, 0.97)***
Named Entities (WHERE)	0.69 (0.52, 0.92)*	0.33 (0.22, 0.48)***	0.97 (0.94, 1.01)	1.07 (1.04, 1.10)***
AI Linguistic Markers				
Affect Word	2.55 (1.83, 3.68)***	0.79 (0.57, 1.09)	0.95 (0.92, 0.98)**	
Past Focus Word	0.46 (0.33, 0.62)***	1.15 (0.79, 1.68)	1.10 (1.06, 1.13)***	1.08 (1.04, 1.11)***
Pronouns	0.60 (0.45, 0.78)***	0.85 (0.62, 1.15)	1.03 (1.00, 1.06)	1.06 (1.03, 1.09)***
Conjunctions	1.97 (1.44, 2.76)***	1.19 (0.88, 1.61)	1.03 (1.00, 1.06)	1.03 (1.00, 1.06)*
Human Heuristics				
Word Count	0.91 (0.70, 1.18)	1.50 (1.07, 2.14)*	0.90 (0.87, 0.93)***	0.65 (0.63, 0.67)***
Proper Nouns			0.93 (0.88, 0.98)*	0.89 (0.85, 0.93)***
Nominal Subjects			0.97 (0.94, 1.01)	1.04 (1.01, 1.07)*
Causation Word			0.96 (0.93, 0.99)**	0.93 (0.90, 0.95)***
Prepositions			0.94 (0.91, 0.97)***	0.94 (0.91, 0.96)***
Quantifiers			0.95 (0.92, 0.98)**	0.95 (0.92, 0.98)***
Adverb			0.96 (0.93, 1.00)*	
Differentiation Word			0.95 (0.92, 0.98)***	
Article Presence			1.04 (1.01, 1.08)**	0.96 (0.93, 0.98)**
CLIP Similarity × Article			1.03 (1.00, 1.06)	1.10 (1.07, 1.14)***
Model Fit				
Constant	2.75 (2.05, 3.76)***	0.83 (0.60, 1.13)	3.11 (2.51, 3.85)***	0.87 (0.76, 1.00)
Observations	408	272	25296	26112
Log-Likelihood	-165.87	-129.56	-13298.54	-16157.41
AIC	349.75	277.12	26635.07	32346.82

with article context over original journalistic captions and struggled to distinguish between journalistic and AI-generated captions. These observations align with previous research concluding that "human heuristics for AI-generated language are flawed," especially in cross-modal image captioning tasks [5].

Finally, comparing Models 3 (AI detection) and 4 (preference judgment), cross-modal features, such as CLIP Similarity, showed stronger associations with caption preference than AI detection. CLIP Similarity was associated with AI detection (OR = 0.96, 95% CI: [0.93, 0.99]; $p < .01$) but more strongly with preference judgments (OR = 0.79, 95% CI: [0.77, 0.82]; $p < .001$). Similarly, its interaction with article presence was linked to AI detection (OR = 1.03, 95% CI: [1.00, 1.06]; $p = .05$) but showed a stronger association with preference judgments (OR = 1.10, 95% CI: [1.07, 1.14]; $p < .001$). These findings suggest that humans prioritize context and image-based cues differently depending on the task. When evaluating caption quality, they appear to rely more on cross-modal relationships, such as how well the caption aligns with the image or article (captured by the CLIP Similarity × Article interaction). On the other hand, when making AI detection judgments, they primarily focus on linguistic cues, rather than cross-modal consistency.

5 Discussion

The development of multimodal generative AI has significantly simplified the process for content creators to generate text from images that closely resembles human-generated content. However, model-based approaches to evaluating image caption quality face limitations, as they often focus on singular evaluation dimensions, whereas human judgment is more flexible, influenced by context and task-specific factors.

Using a naturalistic news image dataset, this study examines how human assessments of AI-generated captions differ when captions are produced with or without contextual information, as measured through caption preference and AI detection tasks. The results highlight that the current LVLMs can perform effective visual reasoning and generate captions that approximate journalist-authored captions. When provided with contextual information, these models can even produce captions that are sometimes preferred over original journalistic captions.

Human assessment of captions can be decomposed into two key factors related to context-sensitive captioning quality: informational cues and linguistic cues. Computational feature extraction and logistic regression reveal that these two dimensions operate differently. Access to article content enhances caption quality by improving name recognition, location specificity, and image-caption semantic alignment—conceptualized as informational signals. However, AI-generated text still retains inherent linguistic patterns that persist even when models are provided with full article context, which we define as AI linguistic markers.

Interestingly, while humans are more likely to correctly interpret informational cues, their intuition about AI linguistic style often diverges from actual AI markers. Lastly, human users are more likely to integrate multiple modalities in preference tasks but rely heavily on linguistic-based strategies for AI detection, demonstrating the fluid and adaptable nature of human judgment in evaluating image-text alignment.

These findings have broader implications for cross-modal reasoning in modern generative AI. While AI-generated text increasingly mirrors human language and is often difficult to distinguish from human-authored content, its true informativeness may still fall short of expert-authored content, and these informative signals might be what humans can reliably depend on.

5.1 Limitations and Future Work

The current work uses only the VOA news dataset. Future research could explore generalizability by extending the analysis to other mainstream news outlets and different types of content, such as Wikipedia articles [6]. Additionally, this study combines observational and experimental methods in a way that deviates from standard experimental procedures. This approach may limit the ability to establish causality, as observed differences in human judgments may be confounded by other factors, such as image complexity or caption length. Furthermore, the operationalization of informational and linguistic features in our current work relies on stepwise feature selection and logistic regression rather than explicitly and gradually manipulating them, which might compromise the robustness of the results.

Our future work can build on this design by incorporating narrower and mismatched contexts to more precisely modulate caption informativeness and directly test whether insensitivity to AI linguistic markers persists under these conditions. Through this subsequent experimental design, we aim to strengthen the replicability and robustness of our findings. Lastly, recognizing that generative AI is highly sensitive to prompt wording, future replications will also explore variations in prompt phrasing when generating captions for both ground truth and human evaluation.

6 Acknowledgments

We are grateful for the support of the OpenAI Research Access Program, which provided access to the resources that enabled the multi-agent simulation in our study. We also appreciate the valuable feedback from the principal investigator and the students in the Computation and Language for Society (Coalas) Lab and the Communicative Mind (Co-Mind) Lab.

References

[1] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2017. Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures (Extended Abstract). In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 4970–4974. doi:10.24963/ijcai.2017/704

[2] G. S. Cheema, S. Hakimov, E. Müller-Budack, C. Otto, J. A. Bateman, and R. Ewerth. 2023. Understanding image-text relations and news values for multimodal news analysis. *Frontiers in Artificial Intelligence* 6 (2023). doi:10.3389/frai.2023.1125533

[3] Z. Gan, L. Li, C. Li, L. Wang, Z. Liu, and J. Gao. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision* 14 (2022), 163–352. doi:10.1561/0600000105

[4] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, and Y. Choi. 2021. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 7514–7528. https://aclanthology.org/2021.emnlp-main.595

[5] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. 2023. Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences* 120, 11 (2023), e2208839120. doi:10.1073/pnas.2208839120

[6] Elisa Kreiss, Cynthia Bennett, Shayan Hooshmand, Eric Zelikman, Meredith Ringel Morris, and Christopher Potts. 2022. Context Matters for Image Descriptions for Accessibility: Challenges for Referenceless Evaluation Metrics. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4685–4697. doi:10.18653/v1/2022.emnlp-main.309

[7] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media Structured Common Space for Multimedia Event Extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 2557–2568. doi:10.18653/v1/2020.acl-main.230

[8] F. Liu, T. Guan, Z. Li, L. Chen, Y. Yacoob, D. Manocha, and T. Zhou. 2023. Hallusionbench: You see what you think? or you think what you see? an image-context reasoning benchmark challenging for gpt-4v (ision), llava-1.5, and other multimodality models. *arXiv preprint arXiv:2310.14566* (2023).

[9] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual News: Benchmark and Challenges in News Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6761–6771. doi:10.18653/v1/2021.emnlp-main.542

[10] Annika Muehlbradt and Shaun K Kane. 2022. What's in an ALT Tag? Exploring Caption Content Priorities through Collaborative Captioning. *ACM Transactions on Accessible Computing (TACCESS)* 15, 1 (2022), 1–32.

[11] Nandita Shankar Naik, Christopher Potts, and Elisa Kreiss. 2024. CommVQA: Situating Visual Question Answering in Communicative Contexts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 13362–13377. doi:10.18653/v1/2024.emnlp-main.741

[12] Habiba Sarhan and Simon Hegelich. 2023. Understanding and evaluating harms of AI-generated image captions in political images. *Frontiers in Political Science* (2023). https://api.semanticscholar.org/CorpusID:262191969

[13] Andrew Taylor Scott, Lothar D Narins, Anagha Kulkarni, Mar Castanon, Benjamin Kao, Shasta Ihorn, Yue-Ting Siu, and Ilmi Yoon. 2023. Improved Image Caption Rating – Datasets, Game, and Model. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 172, 7 pages. doi:10.1145/3544549.3585632

[14] Abigale Stangl, Nitin Verma, Kenneth R. Fleischmann, Meredith Ringel Morris, and Danna Gurari. 2021. Going Beyond One-Size-Fits-All Image Descriptions to Satisfy the Information Wants of People Who are Blind or Have Low Vision. In *Proceedings of the 23rd International ACM SIGACCESS Conference on Computers and Accessibility* (Virtual Event, USA) *(ASSETS '21)*. Association for Computing Machinery, New York, NY, USA, Article 16, 15 pages. doi:10.1145/3441852.3471233

[15] Liming Xu, Quan Tang, Jiancheng Lv, Bochuan Zheng, Xianhua Zeng, and Weisheng Li. 2023. Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing* 546 (2023), 126287. doi:10.1016/j.neucom.2023.126287

[16] Xuewen Yang, Svebor Karaman, Joel Tetreault, and Alejandro Jaimes. 2021. Journalistic Guidelines Aware News Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican Republic,

5162–5175. doi:10.18653/v1/2021.emnlp-main.419

[17] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

[18] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 436, 20 pages. doi:10.1145/3544548.3581318

[19] Mingyang Zhou, Grace Luo, Anna Rohrbach, and Zhou Yu. 2022. Focus! Relevant and Sufficient Context Selection for News Image Captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 6078–6088. doi:10.18653/v1/2022.findings-emnlp.450