Understanding Human Heuristics in Context-Sensitive Image Captioning

Yanru Jiang[1,2], **Rick Dale**[1], **Hongjing Lu**[2,3]
[1] Department of Communication, University of California, Los Angeles
[2] Department of Statistics, University of California, Los Angeles
[3] Department of Psychology, University of California, Los Angeles
{yanrujiang, rdale, hongjing}@ucla.edu

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 resembles human language, its informativeness and coherence, derived from cross-modal image-text reasoning, may still fall short of content generated by human experts. 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 logistic regression analyses reveal that while humans effectively interpret informational cues, their intuition about AI linguistic style often diverges from actual AI language 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.

Keywords: Generative AI, Multimodal Communication, Vision-Language Models, Human-Computer Interaction (HCI), Image Captioning, Model Evaluation

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 (Bernardi et al., 2017). Beyond these technical challenges, recent studies have emphasized that image captioning is highly context-sensitive. These studies reported that the context in which an image appears can significantly influence both the expected informativeness and linguistic style of a caption of the image (Qu, Tuytelaars, & Moens, 2024; Kreiss et al., 2022).

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 (Jakesch, Hancock, & Naaman, 2023; J. Zhou, Zhang, Luo, Parker, & De Choudhury, 2023), while human heuristics for detecting such stylistic differences are often unreliable (Jakesch et al., 2023). 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 (Yang, Karaman, Tetreault, & Jaimes, 2021).

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 (Jakesch et al., 2023; Yang et al., 2021) 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.

Image Captioning

Automatic image description is a highly challenging task that requires machines to perceive and recognize various visual elements (such as objects, actions, and scenes), understand their compositions and semantic relationships, and generate linguistically coherent descriptions that align with human cognition (Xu et al., 2023; Bernardi et al., 2017). 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 (Bernardi et al., 2017).

Previous research highlights that automatic image description and captioning is not a one-size-fits-all task (Naik, Potts, & Kreiss, 2024); 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 (Stangl, Verma, Fleischmann, Morris, & Gurari, 2021). 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.

Evaluating Image Captions

Recent advancements in LVLMs, such as GPT-4V (OpenAI et al., 2024) and Gemini (Google et al., 2024), have demonstrated impressive capabilities in both visual reasoning and language generation. These models can flexibly adapt to different linguistic styles through prompting, mimicking jour-

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

nalistic, conversational, or descriptive tones with minimal effort (Jakesch et al., 2023; Sarhan & Hegelich, 2023). While recent work has begun to explore complex image-text reasoning in captioning and related multimodal tasks (Wan, Cho, Stengel-Eskin, & Bansal, 2024; K. Zhou, Lee, Misu, & Wang, 2024), fine-grained visual understanding and reasoning in LVLMs remains an open challenge.

A common computational approach to evaluating automatic image descriptions and captions involves reference-less metrics, such as CLIPScore (Hessel et al., 2021), which assesses image-caption similarity using pre-trained vision-language models without requiring ground-truth labels (Scott et al., 2023). While these metrics offer efficiency and scalability, they are typically optimized for strict image–caption alignment, and often fail to explicitly account for informational appropriateness, linguistic preference, or context sensitivity (Kreiss et al., 2022).

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 (M. Zhou, Luo, Rohrbach, & Yu, 2022), 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 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.

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 (Cheema et al., 2023). Recent advances in generative AI (Gan et al., 2022) have enabled the creation of AI-generated captions that closely resemble journalist-written ones (Liu et al., 2023), 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 (Sarhan & Hegelich, 2023). 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) (Yang et al., 2021). Thus, access to high-quality contextual information serves as a key factor in the informativeness of AI-generated captions. Previous 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" (Yang et al., 2021).

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 score and the presence of named entities—specifically mentions of "who" and "where," following journalistic conventions (Sarhan & Hegelich, 2023; Yang et al., 2021). Linguistic style, on the other hand, is defined by a set of AI language markers identified in prior AI detection studies (Jakesch et al., 2023). 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 feature 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 associations, 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.

AI-generated News Image Captions

Sampling Image, Caption, Article Triplets

Image-caption stimuli for both 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. (2020). This VOA dataset contains 1,014 news images along with 199 accompanying articles. Each article contains one to eight images, and each image comes with its own original caption written by VOA journalists. To construct the ⟨*image, caption, article*⟩ stimuli for the *with-article* vs. *without-article* conditions, one image-caption pair was randomly sampled from each of the 199 articles. A total of 191 triplets were selected for AI captioning, with balanced topic coverage and exclusion of extreme images (e.g., violence or dead bodies) to comply with IRB requirements.

Control for Caption Concreteness To address a potential confound in the *without-article* condition—namely, that vision-language models often struggle to identify specific people or locations from images alone—we annotated whether both the original and AI-generated captions included concrete information (e.g., names of figures or locations). This step aimed to control for low-level cues participants might use when selecting between captions, such as recognizing named entities that only appear in one caption.

Based on these binary annotations, we identified a sub-

set of 136 "concrete" stimuli from the previously sampled 191 ⟨*image, caption, article*⟩ triplets. A triplet was included in this subset if either both original and AI-generated captions contained concrete information or neither did, ensuring matched concreteness. Pairs where only one caption included concrete details (i.e., imbalanced concreteness) were excluded. This issue does not arise in the *with-article* condition, where the accompanying text typically provides the relevant named entities to the model.

Generating AI Captions

The AI-generated captions for the *without-article* condition were generated using GPT-4V (OpenAI et al., 2024) via the "gpt-4-vision-preview" API, given only the corresponding image. For the *with-article* condition, captions were produced by the same model using both the corresponding image and the associated news articles from 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. All AI captions were generated using default settings. The AI captions were generated over multiple days in November 2023 and February 2024 due to API usage limits. We did not observe substantial changes in caption quality or style over time.

Human Evaluation of Image Captions

Using 136 stimuli, 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 #168450, February 2024).

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 at UCLA. 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.

Procedure

Each experiment employed a between-subjects design to examine how participants select captions based on images, with or without contextual information. In the *with-article* condition, participants saw both the image and article; in the *without-article* condition, only the image was shown (see Figure 1). In both conditions, participants completed two-alternative forced choice (2AFC) tasks across 136 stimulus sets, choosing the better of two captions—one written by a journalist (serving as the baseline) and one generated by AI.

Before the main task, participants reviewed instructions and completed two practice trials. Three attention check trials, each pairing a clearly irrelevant caption with the original human-written one, were randomly interspersed. Participants who failed more than one attention check (i.e., passed fewer than two) were excluded from analysis.

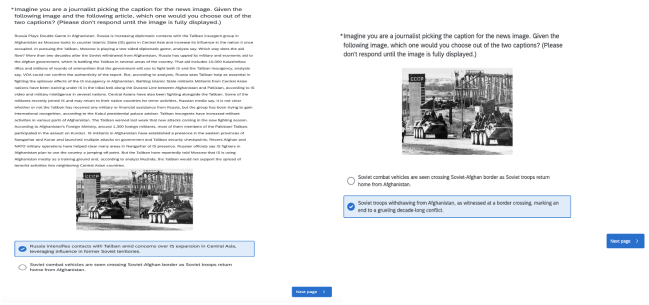

Figure 1: *With-Article* (left) vs. *Without-Article* (right) conditions. Images adapted from Li et al. (2020).

Experiment 1 Human Preference

Experiment 1 compared human preferences for choosing AI-generated captions when articles are provided 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 the Experiment 1 analysis.

Experiment 2 Detecting AI-generated Captions

Experiment 2 aims to assess whether the participants can distinguish AI-generated captions from the journalist-generated captions, both with and without the article being provided (same settings as Experiment 1). Specifically, the insruction was "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 the Experiment 2 analysis.

Computational Feature Extraction

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. (2023).

Linguistic features were adapted from Jakesch et al. (2023)'s machine learning–based feature selection and are relevant to caption generation. These included *Word Count*, *Proper Noun*, and several LIWC-derived categories (Pennebaker, Francis, & Booth, 2001), such as *Affect*, *Past Tense*, *Pronouns*, *Conjunctions*, *Causation*, *Differentiation*, *Quantifiers*, and *Adverbs*.

Informational features focused on visual-semantic integration. We used *CLIPScore* to measure image-caption semantic similarity. Additionally, following the journalistic ⟨who, when, where, what⟩ convention (Yang et al., 2021), we applied spaCy's (https://spacy.io/) named entity recognition (NER) to identify *WHO* entities (PERSON, NORP, ORG) and *WHERE* entities (FAC, GPE, LOC). These entity types, shown to benefit from article context (M. Zhou et al., 2022; Liu, Wang, Wang, & Ordonez, 2021), are also more visually grounded than WHEN or WHAT, making them strong indicators of cross-modal integration.

All extracted features (except *CLIPScore*) were normalized by caption length to control for verbosity across captions. For the subsequent logistic regression analyses, all features were z-score standardized. Given the high dimensionality, we used bidirectional stepwise selection via R's `step()` function to optimize model fit based on AIC, removing weak predictors.

Results

Human vs. Model Judgment

To compare human and model assessments at an aggregated level, we report subject-level caption preference and human-likeness (i.e., reverse-coded AI detection), 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).

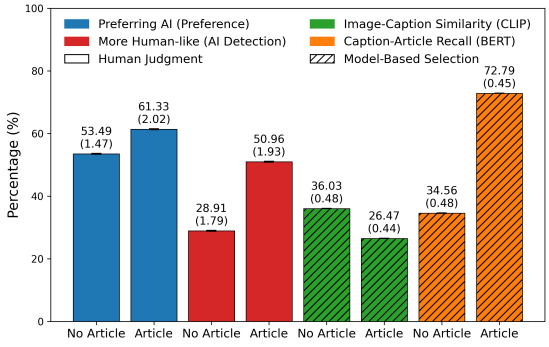

Figure 2: Human vs. Model-Based Judgments Across Conditions. Shown are percentages for AI Caption Preference, Perceived Human-Likeness (1 - AI Detection), Image-Caption Similarity (CLIPScore), and Caption-Article Recall (BERT Recall), across No Article vs. Article conditions.

Human Judgment Human responses were averaged across 136 stimuli, reflecting the proportion of participants selecting AI-generated captions in Experiment 1 (preference) and Experiment 2 (perceived as AI). To align the directional impact of article access across judgments, human-likeness was computed by reverse coding the "detected as AI" responses in Experiment 2.

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

Experiment 2 confirms that participants cannot distinguish AI-generated from human captions when context is provided to AI ($\mu = 50.96\%$, $SD = 1.93\%$). When no article is provided to either the AI model or participants, people can identify AI-generated content (i.e, the error rate of perceiving the AI caption as journalist-generated was 28.91%; $SD = 1.79\%$).

Model-Based Selection Examining model-based assessments reveals important differences from human judgment. For instance, CLIPScore-based judgments show limited alignment with human preferences under both the no-article condition ($\mu = 36.03\%$, $SD = 0.48\%$) and the article condition ($\mu = 26.47\%$, $SD = 0.44\%$). These results suggest that metrics focused solely on image–text semantic alignment may overlook dimensions that human evaluators consider, such as informativeness, relevance, and contextual fit. In fact, the presence of an article—which supports the inclusion of external information in captions—can lower CLIPScore, even though such enrichment is often valued in human assessments of visual storytelling.

While CLIPScore emphasizes visual-semantic alignment, BERT Recall Score (Zhang, Kishore, Wu, Weinberger, & Artzi, 2020) 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 to add the irrelevant information to the image.

These distinctions highlight a divergence between model-based and human-centered evaluations in news image captioning. Rather than indicating shortcomings in existing metrics—which were designed for more narrowly defined alignment tasks—these differences underscore the need for complementary evaluation criteria that better reflect human preferences for informativeness and contextual appropriateness.

Human Heuristics for Generative Content

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 versus journalist-written (Model 1, Table 1, DV: `is AI-generated = 1`). Across all captions, both informational features, WHO entities (OR = 0.48, 95% CI: [0.34, 0.67]) and WHERE entities (OR = 0.69, 95% CI: [0.52, 0.92]), and several linguistic markers were significant predictors (all $p < .001$, except WHERE entities: $p < .05$). These linguistic markers, termed *AI Linguistic Markers*, included Affect Words (OR = 2.55), Past-Focused Words (OR = 0.46), Pronouns (OR = 0.60), and Conjunctions (OR = 1.97).

Next, we ran a separate regression focusing exclusively on AI-generated captions to examine the effect of article access (Model 2, DV: `with article access = 1`). Informational features were significantly associated with captions generated with article access, including CLIP Similarity (OR = 1.67), and WHO (OR = 0.24) and WHERE entities (OR = 0.33; all $p < .001$). In contrast, none of the linguistic markers showed significant differences by article access (all $p > .05$). These findings suggest that the presence of article context is associated with differences in informational content, but not with variation in linguistic style, among 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. In contrast, features that are nonsignificant in Model 2 but significant in Model 1 reinforce the validity of 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.

Informational and Linguistic Cues in Human Evaluation After categorizing features into informational and AI linguistic markers based on their associations with AI-generated captions, we conducted mixed-effects logistic regression with stepwise feature selection and an interaction term (article presence × informational features). This analysis aimed to predict AI detection (Model 4) and human preference (Model 3), both using the dependent variable `human response choosing AI = 1`. Random effects for participants were included to account for individual variability. To align preference judgments with AI detection directionality, we reverse-coded the caption preference scale to reflect "less preferred" captions.

Comparing Model 1 (actual AI-generated captions) with Model 3 (captions perceived as AI-generated), all informational features showed consistent directional associations (WHO: OR = 0.88, 95% CI: [0.84, 0.92]; WHERE: OR = 0.97, 95% CI: [0.94, 1.01]), suggesting humans rely on similar cues to those statistically linked to AI generation. However, linguistic cues showed weaker or reversed associations in Model 3 relative to Model 1. Participants also relied on additional linguistic patterns not predictive of AI generation, which we term *Human Heuristic* features, referring to intuitive but often inaccurate assumptions about AI language.

A similar pattern emerged when comparing Model 1 with Model 4 (captions rated as less preferred): both cross-modal and person-entity features aligned in direction (CLIP Similarity: OR = 0.79, 95% CI: [0.77, 0.82]; WHO: OR = 0.93, 95% CI: [0.89, 0.97]). These findings indicate that while human judgments about caption quality incorporate reliable informational cues, they often diverge from the actual linguistic characteristics of AI-generated text, highlighting discrepancies between human intuition and AI linguistic markers.

These discrepancies may explain why participants preferred AI-generated captions with article access over original journalist-written ones and had difficulty distinguishing between them. These findings align with prior research showing that human heuristics for detecting AI language are limited (Jakesch et al., 2023) and extend to image-to-text tasks.

Finally, comparing Models 3 and 4, cross-modal features like CLIP Similarity were more predictive of preference than AI detection (AI detection: OR = 0.96, 95% CI: [0.93, 0.99]; preference: OR = 0.79, 95% CI: [0.77, 0.82]). Similarly, its interaction with article access was modestly associated with AI detection (OR = 1.03, 95% CI: [1.00, 1.06]; $p > .05$) but more strongly linked to preference (OR = 1.10, 95% CI: [1.07, 1.14]; $p < .001$). These results suggest that humans prioritize cross-modal consistency (such as image-caption and caption-article alignment) when making preference judgments, but primarily rely on linguistic cues and are less sensitive to alignment when making AI detection judgments.

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 are typically designed for singular and strict alignment tasks, 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 demonstrate that current vision-language models are capable of effective visual reasoning, producing captions that closely approximate those written by journalists. 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 qual-

Table 1: Odds Ratios for Context-Related and Context-Independent Features. Sig. levels: * < 0.05, ** < 0.01, *** < 0.001. Odds ratios (OR) and 95% confidence intervals (CIs) were computed by exponentiating the logistic regression coefficients and their bounds: OR $= e^{\beta}$ and 95% CI = $[e^{\beta - 1.96 \cdot SE}, e^{\beta + 1.96 \cdot SE}]$. OR > 1 indicates a positive association with the outcome (e.g., OR = 1.05 suggests a 5% higher odds of the caption being associated with the outcome); vice versa for OR < 1.

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 $\times$ 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

ity: 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 resembles human natural language and is often difficult to distinguish from human-authored content, its informativeness may still fall short of expert-authored content. As multimodal AI systems become more persuasive, these informative signals may serve as a crucial anchor that humans can reliably depend on to discern credibility and resist misinformation.

Limitations and Future Work

This study focuses on the VOA news dataset; future work could test generalizability across other news sources and content types. Secondly, our study manipulated caption informativeness via article access. While this approach makes informativeness a trackable feature, previous studies have used narrower contexts (e.g., news paragraphs with high named-entity coverage related to the image) to further enhance caption quality (M. Zhou et al., 2022). Our future work can build on the current design by incorporating narrowed and mismatched contexts to more precisely alter caption informativeness and examine 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 sensitive to prompt wording, future replications should explore variations in prompt phrasing when generating captions for both ground truth and human evaluation.

References

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., ... Plank, B. (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* (pp. 4970–4974). doi: 10.24963/ijcai.2017/704

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

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

Google, G. T., Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., ... Vinyals, O. (2024). *Gemini: A family of highly capable multimodal models.* Retrieved from https://arxiv.org/abs/2312.11805

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, *120*(11), e2208839120. doi: 10.1073/pnas.2208839120

Kreiss, E., Bennett, C., Hooshmand, S., Zelikman, E., Ringel Morris, M., & Potts, C. (2022, December). Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 4685–4697). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.309

Li, M., Zareian, A., Zeng, Q., Whitehead, S., Lu, D., Ji, H., & Chang, S.-F. (2020, July). Cross-media structured common space for multimedia event extraction. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 2557–2568). Online: Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.230

Liu, F., Guan, T., Li, Z., Chen, L., Yacoob, Y., Manocha, D., & Zhou, T. (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 multi-modality models. *arXiv preprint arXiv:2310.14566*.

Liu, F., Wang, Y., Wang, T., & Ordonez, V. (2021, November). Visual news: Benchmark and challenges in news image captioning. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 6761–6771). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.542/ doi: 10.18653/v1/2021.emnlp-main.542

Naik, N. S., Potts, C., & Kreiss, E. (2024, November). CommVQA: Situating visual question answering in communicative contexts. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 13362–13377). Miami, Florida, USA: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.emnlp-main.741/ doi: 10.18653/v1/2024.emnlp-main.741

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., ... Zoph, B. (2024). *Gpt-4 technical report.* Retrieved from https://arxiv.org/abs/2303.08774

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, *71*(2001), 2001.

Qu, T., Tuytelaars, T., & Moens, M.-F. (2024, June). Visually-aware context modeling for news image captioning. In K. Duh, H. Gomez, & S. Bethard (Eds.), *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)* (pp. 2927–2943). Mexico City, Mexico: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.naacl-long.162/ doi: 10.18653/v1/2024.naacl-long.162

Sarhan, H., & Hegelich, S. (2023). Understanding and evaluating harms of ai-generated image captions in political images. *Frontiers in Political Science*. Retrieved from https://api.semanticscholar.org/CorpusID:262191969

Scott, A. T., Narins, L. D., Kulkarni, A., Castanon, M., Kao, B., Ihorn, S., ... Yoon, I. (2023). Improved image caption rating – datasets, game, and model. In *Extended abstracts of the 2023 chi conference on human factors in computing systems.* New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3544549.3585632

Stangl, A., Verma, N., Fleischmann, K. R., Morris, M. R., & Gurari, D. (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.* New York, NY, USA: Association for Computing Machinery. doi: 10.1145/3441852.3471233

Wan, D., Cho, J., Stengel-Eskin, E., & Bansal, M. (2024). Contrastive region guidance: Improving grounding innbsp;vision-language models without training. In *Computer vision – eccv 2024: 18th european conference, milan, italy, september 29–october 4, 2024, proceedings, part lxxix* (p. 198–215). Berlin, Heidelberg: Springer-Verlag. Retrieved from https://doi.org/10.1007/978-3-031-72986-7$_1$2 doi: 10.1007/978-3-031-72986-7$_1$2

Xu, L., Tang, Q., Lv, J., Zheng, B., Zeng, X., & Li, W. (2023). Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing*, *546*, 126287. doi: https://doi.org/10.1016/j.neucom.2023.126287

Yang, X., Karaman, S., Tetreault, J., & Jaimes, A. (2021, November). Journalistic guidelines aware news image captioning. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Proceedings of the 2021 conference on empirical methods in natural language processing* (pp. 5162–5175). Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.emnlp-main.419/ doi: 10.18653/v1/2021.emnlp-main.419

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). Bertscore: Evaluating text generation with bert. In *International conference on learning representations.*

Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (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.* New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3544548.3581318 doi: 10.1145/3544548.3581318

Zhou, K., Lee, K., Misu, T., & Wang, X. (2024, August). ViCor: Bridging visual understanding and commonsense reasoning with large language models. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Findings of the association for computational linguistics: Acl 2024* (pp. 10783–10795). Bangkok, Thailand: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2024.findings-acl.640/ doi: 10.18653/v1/2024.findings-acl.640

Zhou, M., Luo, G., Rohrbach, A., & Yu, Z. (2022, December). Focus! relevant and sufficient context selection for news image captioning. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the association for computational linguistics: Emnlp 2022* (pp. 6078–6088). Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.findings-emnlp.450/ doi: 10.18653/v1/2022.findings-emnlp.450