Artificial Intelligence Systems and Its Applications (AISA)
Vol. 1, No. 1, July 2025
http://journals.cognispectra.com/index.php/aisa/index
P-ISSN: XXXX-XXXX, E-ISSN: XXXX-XXXX

How Well Do Vision-Language Models Explain Sarcasm? An Evaluation of Multimodal Explanation Quality for Social Media Posts

***Ikhlasul Amal*[1]*, Annisa Nur Ramadhani*[2]**

[1] Universitas Gadjah Mada, Yogyakarta, Indonesia

[2] Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

Abstract: Sarcasm is a complex communicative phenomenon frequently encountered in social media, where the literal meaning of language sharply contradicts the speaker's true intent, often reinforced by multimodal cues such as incongruent images or memes. While prior research has primarily focused on detecting sarcasm, far less attention has been devoted to generating human-interpretable explanations that clarify why content is sarcastic. This study addresses this gap by systematically evaluating the capabilities of fifteen Vision–Language Models (VLMs) of varying parameter sizes to produce multimodal sarcasm explanations under zero-shot and few-shot learning conditions. Using the publicly available MORE dataset of social media posts annotated with concise human-written explanations, we benchmarked each model's outputs with three widely used evaluation metrics, including ROUGE, BERTScore, and Sentence-BERT, to assess both surface-level overlap and deeper semantic alignment. Our findings reveal that smaller models can rival or even outperform larger architectures in n-gram similarity measures, while embedding-based metrics often yield high scores even when generated explanations contradict the ground truth. These results highlight the limitations of current automatic metrics in reliably capturing the nuanced reasoning underlying sarcasm. Overall, this work demonstrates that model scale does not consistently predict explanation quality and underscores the need for more robust evaluation protocols.

Keywords: Multimodal Sarcasm, Social Media, Vision-Language Models, Zero-Shot Learning, Few-Shot Learning, BERTScore

Article History:

Received: 29 June 2025 Accepted: 20 July 2025 Published: 28 July 2025

Corresponding Author: Ikhlasul Amal, Email: ikhlasulamal@mail.ugm.ac.id

DOI: XXXX.XXXX

1. Introduction

"Thanks for the dinner, nothing beats a cold slice of cardboard." This expression illustrates sarcasm, where the literal meaning of the words contrasts sharply with the speaker's true intent [1]. Sarcasm represents a distinctive form of emotional expression in which individuals convey meaning that contradicts their genuine feelings or intentions, often through ironic or contradictory statements [2]. Such expression relies on subtle cues intonation, context, and shared knowledge that are typically understood intuitively by humans. Managing with multimodal information requires an understanding of the data displayed totally different modalities. For machines, especially in multimodal contexts, distinguishing between literal content and underlying intent remains a formidable challenge. When sarcastic text is combined with an incongruent image or meme, the resulting semantic mismatch becomes difficult to decode using literal interpretation alone. This complexity not only hinders classification but also impedes the generation of reliable and interpretable explanations. As a result, effective sarcasm understanding must go beyond detection it must also provide meaningful explanation to support transparency and accountability in Artificial Intelligence (AI).

Sarcasm is increasingly prevalent across social media platforms like X/Twitter, Instagram, and TikTok, where sharp commentary is often paired with contradictory visuals to enhance humor or critique. Recent studies have shown that such verbal-visual incongruity is not incidental, but central to the communicative intent [3]. Consequently, literal-only models frequently fail to detect the irony, leading to system errors that undermine user trust. While interest in multimodal sarcasm detection has surged since 2022 [4], [5], [6], the majority of research continues to focus on classification, with limited attention to explanation generation. This gap underscores the need for a new research direction that not only identifies sarcasm but also rationalizes it in a form comprehensible to humans.

Despite its importance, most existing research on sarcasm remains limited to detection, often leveraging large-scale, fine-tuned models trained on extensive annotated datasets. However, such conditions rarely reflect real-world deployment scenarios, where labeled data may be scarce or unavailable. In these cases, models must generalize with minimal or no task-specific supervision, making zero-shot and few-shot learning critical. Furthermore, few studies have explored how multiple Vision-Language Models (VLMs) with varying parameter sizes perform under these constraints or examined how robust they are to variations in input length and caption structure both common in social media data. Compounding these issues is the lack of consensus on evaluation standards: although ROUGE, BERTScore, and SentenceBERT are widely used, their ability to capture the semantic richness and irony of sarcastic explanations remains uncertain.

This study aims to address these gaps by conducting a comprehensive evaluation of 15 Vision-Language Models with different parameter scales, using a publicly available Multimodal Sarcasm

Explanation dataset. The experiments are designed to assess both zero-shot and few-shot performance. The following research questions guide our investigation:

- **RQ1:** How do VLMs with different parameter sizes perform in generating multimodal sarcasm explanations under zero-shot and few-shot learning settings?
- **RQ2:** How effective are automatic evaluation metrics ROUGE, BERTScore, and SentenceBERT in capturing the quality and relevance of generated sarcasm explanations?
- **RQ3:** Do larger models with higher parameter counts consistently outperform smaller models in multimodal sarcasm explanation tasks, as reflected by evaluation scores?

To address these questions, our contributions are as follows:

- We present a comprehensive benchmarking study of 15 VLMs with varying parameter scales for the task of multimodal sarcasm explanation, evaluated under both zero-shot and few-shot learning settings.
- We conduct a systematic assessment of automatic evaluation metrics ROUGE, BERTScore, and SentenceBERT, to analyze their effectiveness and limitations in capturing the quality and relevance of generated sarcasm explanations.
- We investigate the relationship between model scale and performance, providing empirical evidence on whether larger models consistently outperform smaller models or whether more efficient architectures remain competitive in low-data scenarios.

The remainder of this paper is organized as follows: **Section 2** reviews related work and discusses the theoretical background of sarcasm explanation and vision-language models. **Section 3** outlines the proposed methodology, including dataset usage, model setup, and evaluation design. **Section 4** presents the experimental results and analysis based on the research questions. Finally, **Section 5** concludes the paper and proposes future research directions for advancing multimodal sarcasm explanation.

2. Related Work

2.1 Multimodal Sarcasm Explanation

The study of sarcasm in computational linguistics has seen a growing interest in multimodal sarcasm explanation, which moves beyond traditional detection tasks by generating rationales that clarify why a given post is sarcastic [7]. Early research mostly concentrated on classification problems, leveraging both textual and visual features to determine whether an input is sarcastic [8].

A major advancement came with the release of the MuSE dataset, introduced by [1]. in their study titled *"Nice perfume. How long did you marinate in it?"*. This dataset uniquely pairs sarcastic memes with corresponding human-written explanations, encouraging models to produce interpretive outputs rather than simple binary classifications. MuSE was evaluated using a Transformer-based model, setting a foundation for future work in sarcasm explanation.

In dialogue-based contexts, Kumar et al. introduced MOSES, a dataset designed for multimodal sarcasm explanation in conversational settings. Their work demonstrated how sarcasm explanations could be extended across multi-turn dialogue, incorporating not only the sarcastic utterance but also its surrounding context. Building upon this, the MAF framework applied cross-modal fusion techniques for sarcasm explanation in multi-party conversations, further enriching the field [9].

Beyond explanation, a number of studies have continued to focus on multimodal sarcasm detection. Chen et al. developed the CS4MSD framework, which integrates CLIP with contrastive sentiment signals to detect sarcasm based on textual-visual incongruity [10], [11]. Similarly, SarcNet, introduced by Yue et al. in 2024, expands the multilingual and multimodal coverage of sarcasm by providing annotations for both text and image modalities independently [12].

A comprehensive review by Farabi et al. in 2024 outlined recent advances and trends in multimodal sarcasm research, highlighting the increasing role of large-scale vision-language models in this area [13]. Other datasets and frameworks have emerged, such as MMSD2.0 [10], which adopts a multi-view strategy using CLIP, and models that incorporate audio signals alongside visual and textual inputs, such as the approach proposed by Wang et al. (2025) [9]. Recent developments have also proposed sentiment-aware and representation-aligned deep learning frameworks that enhance sarcasm detection by modeling cross-modal emotional incongruity and refining multimodal fusion strategies [14].

These studies collectively illustrate the field's shift toward deeper and more interpretable sarcasm understanding through multimodal architectures, benchmark datasets, and hybrid input modalities.

2.2 VLM for Sarcasm Detection

In recent years, VLMs have proven highly effective in handling sarcasm detection tasks that combine both text and image modalities. A notable advancement was the introduction of S^3 Agent, described by Wang et al. in 2024. They proposed a multi-view agent framework based on large VLMs for zero-shot multi-modal sarcasm detection, achieving a 13.2% accuracy improvement on the MMSD2.0 dataset through perspectives of superficial expression, semantic information, and sentiment alignment [15].

Further research by Tang et al. (2024) leveraged generative Large Language Models (LLMs) with visual instruction and demonstration retrieval techniques. Their model achieved state-of-the-art performance on both in-domain (MMSD2.0) and out-of-domain (RedEval) test sets, highlighting the importance of LLM-based prompting and visual context in sarcasm detection [16].

More recently, Zhang et al. (2025) introduced Commander-GPT, a multi-modal chain-of-thought framework that decomposes the sarcasm detection task into sub-tasks handled by specialized agents.

This method yielded an impressive 19.3% F1 improvement across MMSD benchmarks without any task-specific fine-tuning [17].

Similarly, Ramakrishnan et al. (2025) presented IRONIC, which builds coherence-aware reasoning chains in a zero-shot setting. By leveraging inter-modal coherence relations, IRONIC outperformed existing multi-modal baselines, demonstrating the power of reasoning-inspired architectures for sarcasm detection [18].

Complementing these agent-based approaches, it proposed VisLingInstruct, an autonomous instruction optimization framework for multi-modal LLMs. By refining instruction prompts and improving visual feature extraction, VisLingInstruct enhanced zero-shot performance across several standard benchmarks [15].

Collectively, these studies highlight a shift from fusion-focused architectures to agent- and reasoning-based frameworks that exploit VLLMs' inherent reasoning capabilities. This move enhances adaptability in low-resource settings and robustness across various domains an essential evolution for real-world sarcasm detection systems.

3.3 Zero-shot and Few-shot Prompting

Prompting strategies, especially zero-shot and few-shot techniques, have gained prominence as efficient alternatives to full-model fine-tuning in Natural Language Processing (NLP) and Vision-Language tasks [19]. These approaches allow large language models (LLMs) and VLMs to generalize to new tasks by conditioning their outputs on task-descriptive prompts and a few demonstration examples, rather than relying on task-specific training data [20].

The emergence of GPT-3 marked a significant transition in natural language processing, showcasing how large language models could tackle diverse NLP challenges through carefully designed prompts with minimal training examples. This approach subsequently evolved to encompass multimodal applications, where architectures such as Flamingo and GPT-4V integrate visual encoding components with language generation modules to process combined image-text-instruction inputs, enabling sophisticated capabilities including image captioning, visual question answering, and visual commonsense reasoning [21], [22].

Within multimodal sarcasm research, recent investigations by Tang et al. (2024) and Wang et al. (2024) have employed prompt-based methodologies for sarcasm detection using pre-trained vision-language models without additional fine-tuning [23], [24]. These studies revealed that prompt engineering elements such as linguistic cues, reasoning frameworks, and the sequential arrangement of visual and textual components substantially influence model performance. Building on these advancements, researchers have proposed dynamic prompt optimization techniques that adaptively refine instructions to enhance sarcasm detection accuracy, outperforming traditional static prompting methods [25].

Despite these advances, most prompt-based multimodal studies remain detection-focused, rarely exploring explanation generation. Additionally, few studies investigate how model size and prompt design interact in shaping output quality under low-resource conditions [26]. In contrast to prior studies, our work systematically benchmarks a wide range of VLMs on multimodal sarcasm explanation, with attention to model scaling and evaluation metrics.

3. Methods

3.1 Dataset

The dataset used in this study is MORE (*`https://github.com/LCS2-IIITD/ Multimodal-Sarcasm-Explanation-MuSE)`*, which was released in the paper [1]. This dataset comprises 3,510 social media posts where both images and their corresponding captions collaboratively convey sarcastic meaning. Each post is accompanied by a concise human-written explanation sentence that explicitly reveals the underlying irony, making MORE one of the few resources that provide multimodal sarcasm data paired with interpretative explanations. The dataset is divided into training, validation, and test subsets. In this research, we specifically utilize the test subset to benchmark the performance of various models, comprising 352 examples that include the original images, captions, and corresponding ground-truth explanations. Leveraging this dataset allows for a rigorous evaluation of model capabilities in understanding and explaining sarcasm in a multimodal context.

3.2 Exploratory Data Analysis (EDA)

Before conducting model experiments, we performed a targeted exploratory analysis of the MORE test set to characterize the length of the ground truth explanations. For this purpose, we calculated the token length by splitting each explanation string on whitespace. The average explanation length is 12.60 tokens, with most examples containing fewer than 25 tokens. Only a small number of outliers exceed 40 tokens.

This distribution indicates that the majority of reference explanations are concise and can be generated comfortably within a relatively short output window. Based on this analysis, we set a maximum generation length of 50 tokens to ensure that even the longest explanations could be covered without truncation while also preventing models from producing excessively verbose outputs. Figure 1 illustrates the histogram of explanation token lengths in the test set.

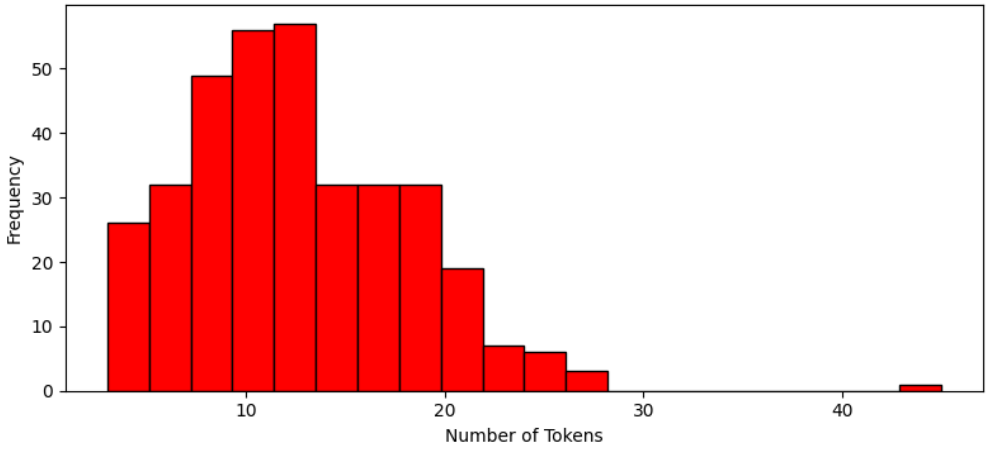


Figure 1: Token Length Distribution of Ground Truth Explanations

3.3 Model and Prompting

In this section, we present our comprehensive evaluation framework for Vision-Language Models (VLMs) ranging from 0.5B to 4B parameters. Our experimental design encompasses model selection, evaluation protocols, computational infrastructure, and hyperparameter exploration to ensure robust and reproducible results. We evaluated 15 VLMs spanning different architectural paradigms and parameter scales. The selected models represent diverse approaches to vision-language understanding, enabling comprehensive analysis across various model families:

Below One Billion Parameters:

- **llava-hf/llava-onevision-qwen2-0.5b-ov-hf**: A compact variant of the LLaVA-OneVision architecture that integrates Qwen2 language backbone with vision encoder for multimodal understanding. This model demonstrates efficient parameter utilization for vision-language tasks while maintaining competitive performance [27].
- **HuggingFaceTB/SmolVLM2-500M-Video-Instruct**: An instruction-tuned variant of the SmolVLM2 architecture specifically designed for video understanding tasks. The model incorporates temporal reasoning capabilities within a compact parameter budget [28].
- **apple/FastVLM-0.5B-Stage3**: Apple's efficient vision-language model optimized for fast inference while maintaining accuracy. The three-stage training approach focuses on progressive capability development from basic vision-text alignment to complex reasoning [29].

One Billion Parameters:

- **ByteDance/Sa2VA-1B**: A streamlined architecture that employs sparse attention mechanisms for efficient vision-language processing. The model demonstrates strong performance on visual question answering while maintaining computational efficiency [30].
- **apple/FastVLM-1.5B-Stage3**: The scaled-up version of FastVLM with enhanced capacity for complex multimodal reasoning tasks. The model benefits from increased parameter count while maintaining the efficient training paradigm [29].
- **deepseek-ai/deepseek-vl-1.3b-chat:** A conversational vision-language model that emphasizes natural dialogue capabilities combined with visual understanding. The architecture incorporates specialized training for multi-turn visual conversations [31].

Two Billion Parameters:

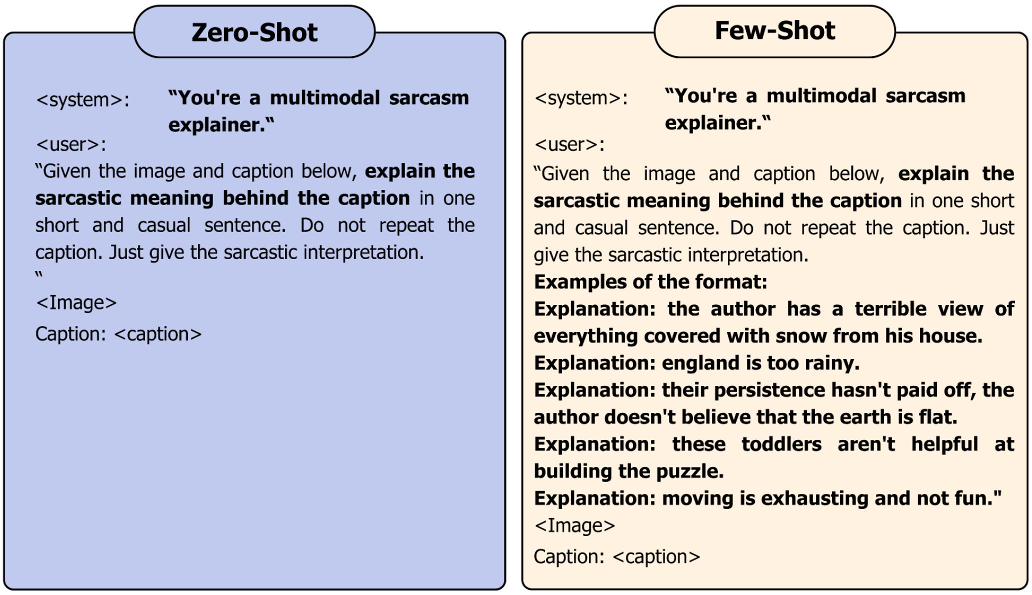
- **HuggingFaceTB/SmolVLM2-2.2B-Instruct**: The instruction-tuned variant of SmolVLM2 that demonstrates strong performance on instruction-following tasks across vision and language modalities. The model architecture emphasizes efficient cross-modal attention mechanisms [28].
- **Qwen/Qwen2-VL-2B-Instruct:** An instruction-tuned vision-language model from the Qwen2 family that integrates advanced visual encoding with powerful language understanding capabilities. The model shows particular strength in detailed visual description and reasoning tasks [32].
- **ibm-granite/granite-vision-3.2-2b:** IBM's granite-based vision-language model that leverages enterprise-grade training data and robust architectural design for reliable multimodal performance across diverse domains [33].

Three Billion Parameters:

- **Qwen/Qwen2.5-VL-3B-Instruct:** The latest iteration of Qwen's vision-language series, incorporating architectural improvements and enhanced training procedures for superior multimodal understanding and instruction following [32].
- **TencentBAC/TBAC-VLR1-3B-preview:** Tencent's vision-language reasoning model that emphasizes logical reasoning capabilities combined with visual understanding, particularly designed for complex multimodal reasoning tasks https://huggingface.co/TencentBAC/TBAC-VLR1-3B-preview [34].
- **deepseek-ai/deepseek-vl2-tiny:** The compact version of DeepSeek's second-generation vision-language architecture, featuring improved efficiency and performance compared to its predecessor while maintaining a small parameter footprint [31].

Four Billion Parameters:

- **microsoft/Phi-3.5-vision-instruct:** Microsoft's Phi-3.5 model extended with vision capabilities, featuring efficient parameter utilization and strong performance on instruction-following tasks across modalities [35].
- **google/gemma-3-4b-it:** Google's instruction-tuned Gemma model extended for vision-language tasks, leveraging the robust Gemma architecture with multimodal capabilities for comprehensive understanding [36].
- **ByteDance/Sa2VA-4B:** The larger variant of the Sa2VA architecture with enhanced capacity for complex vision-language understanding while maintaining the efficient sparse attention mechanisms [30].



Zero-Shot

<system>: **"You're a multimodal sarcasm explainer.‟**
<user>:
"Given the image and caption below, **explain the sarcastic meaning behind the caption** in one short and casual sentence. Do not repeat the caption. Just give the sarcastic interpretation.
‟
<Image>
Caption: <caption>

Few-Shot

<system>: **"You're a multimodal sarcasm explainer.‟**
<user>:
"Given the image and caption below, **explain the sarcastic meaning behind the caption** in one short and casual sentence. Do not repeat the caption. Just give the sarcastic interpretation.
Examples of the format:
Explanation: the author has a terrible view of everything covered with snow from his house.
Explanation: england is too rainy.
Explanation: their persistence hasn't paid off, the author doesn't believe that the earth is flat.
Explanation: these toddlers aren't helpful at building the puzzle.
Explanation: moving is exhausting and not fun."
<Image>
Caption: <caption>

Figure 2: Illustration of prompting setup.

Our evaluation framework employs two primary prompting strategies to assess model capabilities comprehensively. From Figure 2 We designed a consistent chat-style prompt format to evaluate all models in both zero-shot and few-shot settings. The prompt simulates a helpful assistant tasked with explaining the sarcastic meaning of a multimodal social media post. Zero-shot prompting involves providing the model only with the image, the caption, and a single instruction that explicitly asks the model to interpret the sarcasm without repeating the caption. The model has no prior examples and must rely solely on its pretrained knowledge. The instruction is kept concise and clear to avoid ambiguity. Few-shot prompting, by contrast, includes several example explanations before the actual task. These examples serve as demonstrations of the desired output format and tone. Although they are not tied to the current image or caption, they help guide the model toward generating concise, human-like sarcastic interpretations. In our prompt, we provide five short explanation examples.

All experiments were conducted on NVIDIA Tesla T4 GPUs, providing standardized computational resources across all model evaluations. We adopted a consistent decoding configuration across all models to ensure comparability and reduce confounding factors. Specifically, we set *max_new_tokens=50* to accommodate the longest ground truth explanations while preventing overly verbose generations, and used deterministic decoding *do_sample=False* to produce stable outputs for evaluation.

This approach ensures a robust and fair assessment of VLM capabilities across different scales and architectures, offering reliable insights into the current state of vision-language understanding in compact model formats.

3.4 Evaluation Metrics

Our evaluation of multimodal sarcasm explanation capabilities employs three complementary metrics that capture different aspects of text quality and semantic similarity. These metrics provide a comprehensive assessment framework that evaluates both surface-level textual similarity and deeper semantic understanding.

3.4.1 ROUGE Metrics

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) serves as our primary lexical overlap metric, measuring the similarity between generated explanations and reference explanations through n-gram matching [37]. We employ three ROUGE variants to capture different granularities of textual similarity:

- ROUGE-1: Measures unigram overlap between generated and reference texts, providing insights into vocabulary coverage and basic content similarity. This metric evaluates whether the model captures the essential keywords and concepts present in human-written sarcasm explanations. ROUGE-1 is particularly valuable for assessing whether models identify key sarcastic elements, contradictions, or contextual cues that are explicitly mentioned in reference explanations.
- ROUGE-2: Evaluates bigram overlap, capturing local phrase-level similarities and basic syntactic patterns. This metric provides insights into whether models maintain coherent phrase structures and can reproduce common sarcasm-related expressions found in reference explanations. ROUGE-2 is essential for understanding how well models capture the linguistic patterns typical of sarcasm explanation discourse.

- ROUGE-L: Measures the longest common subsequence between generated and reference texts, emphasizing sentence-level structural similarity and maintaining the overall flow of explanation. This metric evaluates whether models preserve the logical progression and argumentative structure present in human explanations of sarcastic content. ROUGE-L is particularly important for sarcasm explanation tasks as it captures the coherence of reasoning chains that explain why content is sarcastic.

3.4.2 BERTScore

BERTScore addresses the limitations of lexical-based metrics by leveraging contextualized embeddings from pre-trained BERT models to evaluate semantic similarity. Unlike ROUGE metrics that rely on exact token matching, BERTScore computes similarity scores based on contextual embeddings, enabling evaluation of semantic equivalence even when surface forms differ [38].

For multimodal sarcasm explanation evaluation, BERTScore provides crucial advantages by capturing semantic relationships that may not be apparent through lexical overlap. Sarcasm explanations often involve paraphrasing, synonymous expressions, and conceptually equivalent statements that describe the same underlying sarcastic mechanism. BERTScore effectively identifies these semantic similarities, providing a more nuanced evaluation of explanation quality.

The metric computes precision, recall, and F1 scores by matching tokens between generated and reference texts based on their BERT embeddings. This approach enables recognition of semantically similar explanations that might use different vocabulary to describe the same sarcastic elements, contextual contradictions, or ironic situations.

3.4.3 Sentence-BERT (SentBERT)

SentBERT extends BERT's capabilities to sentence-level semantic similarity evaluation, providing holistic assessment of explanation quality at the complete utterance level. Unlike token-level matching approaches, SentBERT evaluates the overall semantic coherence and meaning preservation of entire explanation passages [39].

For sarcasm explanation tasks, SentBERT is particularly valuable because it captures the global semantic structure of explanations, including complex relationships between different parts of the reasoning process. Sarcasm explanations typically involve multi-step reasoning that connects visual context, textual content, and implicit social or cultural knowledge. SentBERT's sentence-level evaluation approach effectively captures whether generated explanations maintain the overall logical coherence and semantic integrity of reference explanations.

The metric computes cosine similarity between sentence embeddings of generated and reference explanations, providing a continuous similarity score that reflects the degree of semantic alignment. This approach is especially beneficial for evaluating creative or diverse explanation styles that may be semantically equivalent to references while using different linguistic expressions.

These three metric categories work synergistically to provide comprehensive evaluation coverage. ROUGE metrics ensure that models capture essential lexical elements and maintain structural coherence, BERTScore validates semantic equivalence at the token level, and SentBERT confirms overall meaning preservation at the discourse level. This multi-layered evaluation approach is particularly crucial for sarcasm explanation tasks, where successful performance requires both precise identification of sarcastic elements and coherent articulation of the underlying mechanisms that create sarcastic meaning.

The combination of these metrics enables robust evaluation that accounts for the complexity and diversity inherent in human explanations of sarcastic content, providing reliable assessment of model capabilities across different dimensions of text quality and semantic understanding.

This comprehensive methodology ensures robust evaluation of VLM capabilities across different scales and architectures, providing reliable insights into the current state of vision-language understanding in compact model formats.

4. Results and Discussion

This section presents the results, detailing the actual parameter counts of all models, their comparative performance in zero-shot and few-shot settings, illustrative case studies of metric limitations, and comparisons with prior multimodal baselines.

4.1 Model Parameter Scale Overview

Table 1: Actual Parameter Counts of Evaluated VLMs

Model	Model Number Parameters
FastVLM-0.5B	622,403,552
FastVLM-1.5B	1,675,423,328
SmolVLM-500M	507,482,304
SmolVLM-2.2B	2,246,784,880
Qwen2-VL-2B	2,208,985,600
Qwen2.5-VL-3B	3,754,622,976
DeepSeek-VL-1.3B	1,975,235,584

DeepSeek-VL2-Tiny	3,370,501,440
Sa2VA-1B	1,163,665,458
Sa2VA-4B	3,941,809,458
GraniteVision3.2-2B	2,975,396,928
TBAC-VLR1	3,754,622,976
LLaVA-OV-0.5B	893,675,552
Gemma3-4B	4,300,079,472
Phi3.5-Vision	4,146,621,440

Table 1 presents the precise parameter counts for all evaluated models. While many architectures are labeled according to nominal parameter sizes, the actual number of parameters often diverges substantially from these labels. For instance, DeepSeek-VL-1.3B includes approximately 1.97 billion parameters, significantly exceeding its nominal size, while Sa2VA-4B comprises about 3.94 billion parameters, falling slightly short of the 4-billion mark. This observation reflects the necessity of reporting the actual parameter counts to provide a clearer understanding of each model's scale beyond nominal labels. Figure 3 illustrates the distribution of parameter scales across all evaluated models.

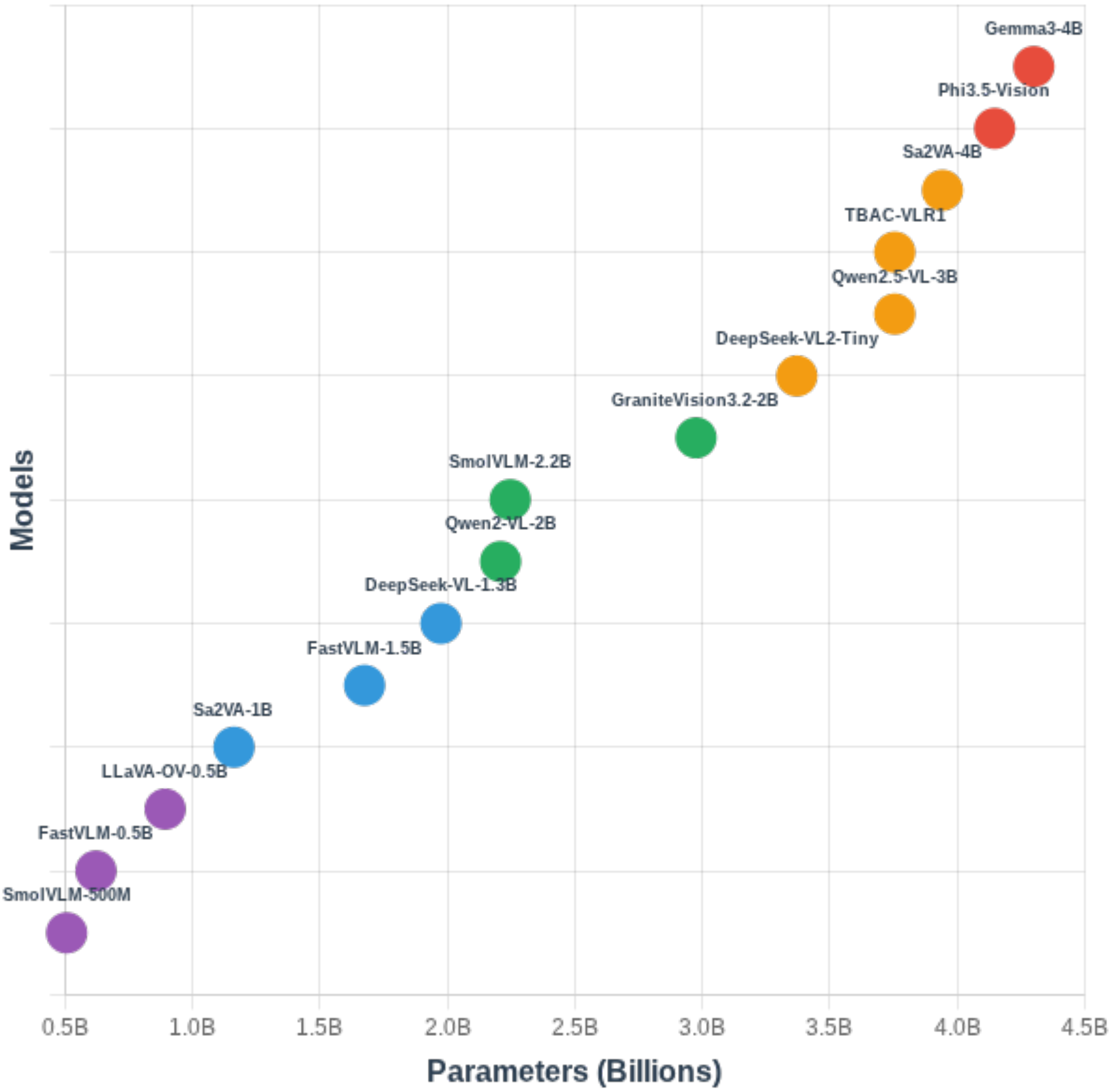


Figure 3: Distribution of Model Parameter Scales

4.2 Zero-shot and Few-shot Performance Comparison

Table 2 and Table 3 reports the comparative performance of all evaluated models in both zero-shot and few-shot settings, measured across five metrics: ROUGE-1, ROUGE-2, ROUGE-L, BERTScore F1, and SentBERT. Overall, the results show substantial variability not only between models of different scales but also within the same model across different evaluation criteria.

Table 2: Zero-shot Performance Across Evaluation Metrics

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1	SentBERT
FastVLM-0.5B	22.40	8.75	19.02	85.67	45.93
FastVLM-1.5B	23.59	8.23	19.44	85.95	47.30
SmolVLM-500M	**31.58**	**13.94**	**27.86**	**88.64**	50.39
SmolVLM-2.2B	25.78	9.25	21.66	87.38	45.23
Qwen2-VL-2B	24.65	7.24	20.65	87.21	45.56
Qwen2.5-VL-3B	22.36	5.02	17.61	86.42	46.48
DeepSeek-VL-1.3B	21.54	8.04	18.43	85.82	42.08
DeepSeek-VL2-Tiny	22.52	6.46	18.51	86.00	42.98
Sa2VA-1B	24.55	7.03	20.60	87.19	47.22
Sa2VA-4B	23.76	6.80	19.20	86.81	46.06
GraniteVision3.2-2B	21.31	5.60	17.53	86.33	45.55
TBAC-VLR1	22.28	5.16	17.58	86.38	46.34
LLaVA-OV-0.5B	27.75	9.14	24.10	87.41	49.55
Gemma3-4B	15.96	1.82	12.64	86.18	39.63
Phi3.5-Vision	24.14	6.92	19.63	86.98	**50.54**

In the zero-shot scenario, SmolVLM-500M achieved the highest ROUGE scores among all models, with ROUGE-1 reaching 31.58 and ROUGE-L reaching 27.86. These values notably exceed those of significantly larger models such as Sa2VA-4B (ROUGE-1: 23.76) and Qwen2.5-VL-3B (ROUGE-1: 22.36), suggesting that smaller models can be highly competitive in producing surface-level n-gram overlaps with reference explanations. A similar pattern emerged in BERTScore F1, where most models clustered closely between 85.5 and 88.6, indicating consistent semantic similarity across architectures. SmolVLM-500M again achieved the highest BERTScore F1 of 88.64, while the lowest was recorded by FastVLM-0.5B at 85.67. SentBERT scores, reflecting sentence-level embedding similarity, exhibited greater dispersion, with Gemma3-4B obtaining the lowest score (39.63) and Phi3.5-Vision reaching the highest (50.54).

In the few-shot condition, several models benefited from prompt-based guidance, with increases particularly visible in the larger architectures. For example, Sa2VA-4B improved its ROUGE-1 from 23.76 to 26.64 and its SentBERT from 46.06 to 49.30. Similarly, Qwen2-VL-2B saw increases across ROUGE metrics and SentBERT similarity, demonstrating responsiveness to few-shot examples. However, improvements were not uniform across the board. SmolVLM-500M, which had been the top performer in ROUGE during zero-shot evaluation, experienced a noticeable drop in ROUGE-1 to 22.97 and a decline in SentBERT similarity to 35.92, suggesting that additional examples may have reduced its generation fidelity or consistency relative to the ground truth.

Table 3: Few-shot Performance Across Evaluation Metrics

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore F1	SentBERT
FastVLM-0.5B	19.34	7.10	16.21	85.58	41.79
FastVLM-1.5B	22.26	7.63	19.10	86.30	45.33
SmolVLM-500M	22.97	8.36	20.35	87.72	35.92
SmolVLM-2.2B	**28.09**	**13.33**	**25.25**	87.51	43.42
Qwen2-VL-2B	25.86	9.15	22.24	87.77	46.20
Qwen2.5-VL-3B	22.34	6.06	18.11	86.60	46.68
DeepSeek-VL-1.3B	20.42	7.46	18.36	**87.98**	32.02
DeepSeek-VL2-Tiny	21.18	8.70	18.18	85.37	29.81
Sa2VA-1B	21.96	6.72	18.72	87.13	40.27
Sa2VA-4B	26.64	8.95	22.53	87.13	49.30
GraniteVision3.2-2B	25.08	8.62	21.30	87.28	46.52
TBAC-VLR1	22.28	5.70	17.98	86.59	46.73
LLaVA-OV-0.5B	26.95	10.19	23.23	87.79	**49.47**
Gemma3-4B	16.93	1.85	14.07	86.66	40.14
Phi3.5-Vision	22.75	8.18	20.02	87.38	40.30

Interestingly, BERTScore F1 remained relatively stable across most configurations, with few-shot performance ranging narrowly from approximately 85.3 to 87.9, showing less sensitivity to the prompt condition than other metrics. This indicates that while the models' token-level similarity varied with prompt inclusion, their overall semantic content alignment was comparatively robust.

Overall, these results highlight that parameter scale and model capacity did not consistently correlate with superior performance across all metrics. Instead, smaller models such as SmolVLM-500M and LLaVA-OV-0.5B frequently matched or outperformed larger models in both zero-shot and few-shot configurations on specific metrics.

The variability in SentBERT and BERTScore F1 further emphasizes the importance of combining n-gram-based and embedding-based evaluations to capture both surface similarity and deeper semantic alignment. To better illustrate these dynamics, Figure 4 and Figure 5 provides a visual comparison of BERTScore F1 and SentBERT scores across all models, highlighting the relative gains and declines between the zero-shot and few-shot settings.

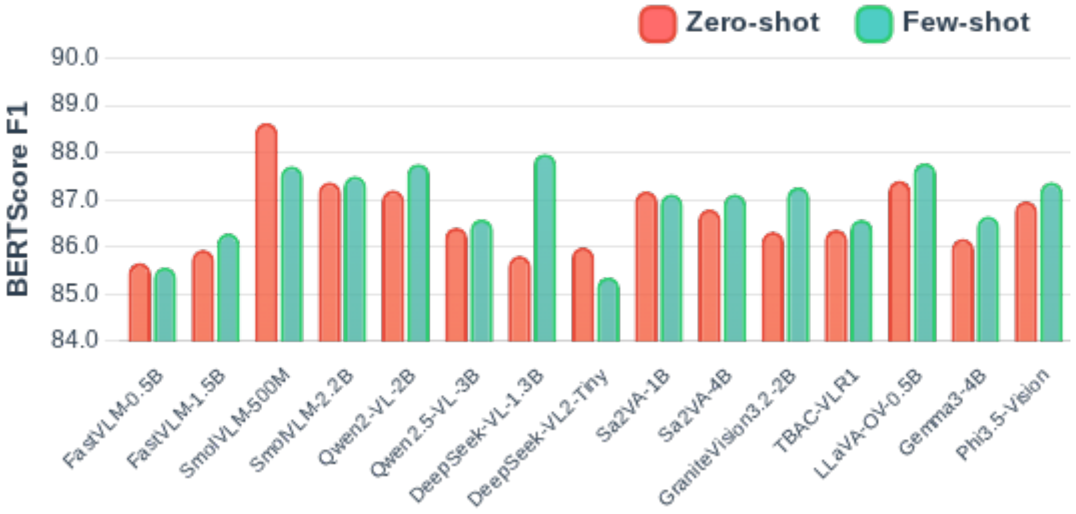


Figure 4: BERTScore F1 Across Models in Zero-shot and Few-shot Settings.

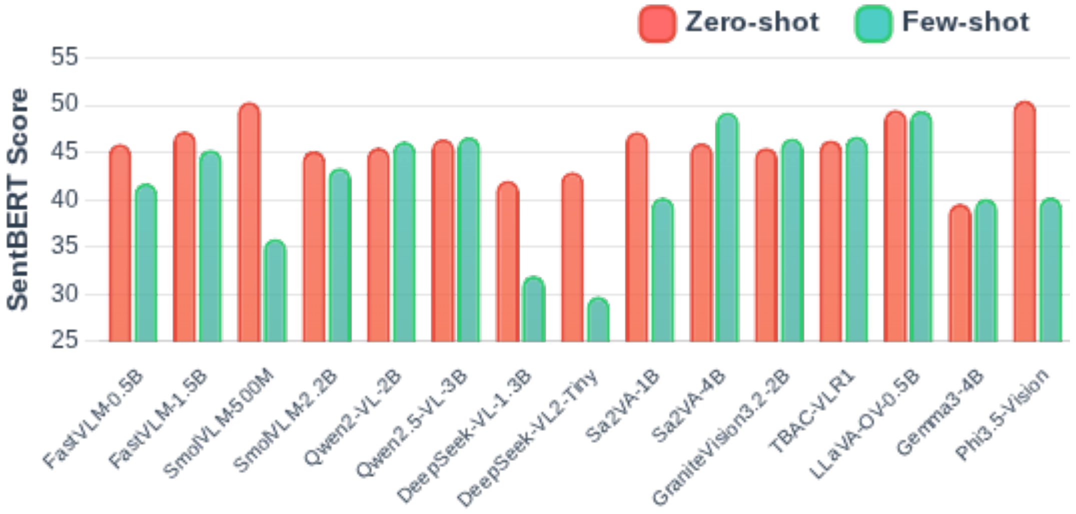


Figure 5: SentBERT Similarity Across Models in Zero-shot and Few-shot Settings.

4.3 Examples Highlighting Metric Divergence

To complement the aggregate metrics reported previously, we present two representative examples that illustrate how different evaluation measures can yield contrasting perspectives on the quality of generated explanations.

Caption: nothing says healthy like chicken nuggets covered in hot sauce ! !

Ground truth: chicken nuggets covered in hot sauce aren't healthy.

Prediction: The caption is sarcastic because chicken nuggets covered in hot sauce are generally considered unhealthy and not a nutritious meal option.

Model: Phi3.5-Vision (Zero-Shot)

Rouge-1	Rouge-2	Rouge-L	BERTScore F1	SentBERT
40.00%	35.71%	40.00%	91.57%	84.26%

Figure 6: Example Highlighting High Embedding Similarity Despite Low ROUGE.

Figure 6 shows an output from Phi3.5-Vision in the zero-shot setting. In this case, although the generated prediction provides a paraphrased explanation of why the caption is sarcastic, the n-gram overlap with the ground truth remains relatively modest, as indicated by ROUGE-1 and ROUGE-L scores (40.00). Nevertheless, BERTScore F1 reached 91.57, suggesting that despite low lexical similarity, the semantic content is highly aligned. This example highlights the importance of incorporating embedding-based metrics, which are more tolerant of paraphrasing and can better capture the intended meaning of the explanation.

The poorest regions in the UK are the poorest in Northern Europe

Caption: # brexit explained see how the eu keeps all the money for itself ? yes

Ground truth: the eu doesn't keep the money for itself.

Prediction: The EU keeps all the money for itself.

Model: SmolVLM-500M (Zero-Shot)

Rouge-1	Rouge-2	Rouge-L	BERTScore F1	SentBERT
82.35%	53.33%	82.35%	92.30%	86.56%

Figure 7: Example Showing Contradictory Predictions with High Similarity Scores.

In contrast, Figure 7 highlights a critical limitation of automatic similarity metrics when evaluating explanation quality. In this example, the prediction simply states, "The EU keeps all the money for itself," whereas the ground truth explicitly negates this claim with, "the EU doesn't keep the money for itself." Despite the fact that the predicted explanation conveys the opposite meaning of the reference, the ROUGE and BERTScore metrics nevertheless report very high similarity scores, with ROUGE-1 and ROUGE-L 82.35 and BERTScore F1 92.30.

This discrepancy arises because these metrics primarily rely on surface-level token overlap and embedding proximity without robust mechanisms to detect negation or factual contradiction. As a result, such outputs can appear highly similar according to automatic measures while effectively conveying misinformation. This example underscores the need for caution when interpreting similarity scores in isolation. Future work could incorporate additional verification steps, such as entailment detection, contradiction analysis, or human evaluation to mitigate the risk of overestimating the quality and factual correctness of generated explanations.

4.4 Comparison with Existing Methods

To contextualize the performance of the evaluated VLMs, we compare their zero-shot and few-shot results against ExMore and ExMoreOCR [1], which were introduced in prior work as baseline methods for this dataset. ExMore employs a multimodal Transformer-based encoder-decoder framework that integrates image and caption features. Specifically, the architecture combines VGG-19 for visual encoding with BART for textual encoding and decoding. Cross-modal learning is performed by projecting captions as queries and image regions as keys and values within a Transformer encoder. The final cross-modal representation is subsequently decoded by a pre-trained BART decoder fine-tuned on the MORE dataset.

In addition to the standard ExMore configuration, ExMoreOCR introduces an OCR text stream as a third modality, processed in parallel with the caption-image encoder. A gating mechanism dynamically weights the relative contributions of image and OCR features before decoding. This tri-modal approach was shown to yield improvements over simpler multimodal baselines in the original work.

Table 4: Comparison with ExMore and ExMoreOCR Baselines

Model	ROUGE-1	ROUGE-2	ROUGE-L	BERT F1-Score	SentBERT
SmolVLM-500M (Zero-Shot)	**31.58**	**13.94**	**27.86**	**88.64**	50.39
Phi3.5-Vision (Zero-Shot)	24.14	6.92	19.63	86.98	50.54
SmolVLM-2.2B (Few-Shot)	28.09	13.33	25.25	87.51	43.42
DeepSeek-VL-1.3B (Few-Shot)	20.42	7.46	18.36	87.98	32.02
LLaVA-OV-0.5B (Few-Shot)	26.95	10.19	23.23	87.79	49.47
ExMore [1]	27.55	12.49	25.23	87.90	59.12
ExMoreOCR [1]	24.23	9.89	22.27	87.00	**59.57**

Table 4 presents the comparison between ExMore, ExMoreOCR, and selected VLMs evaluated in this study. Notably, SmolVLM-500M, despite being a significantly smaller model operating in a zero-shot configuration, achieves the highest ROUGE-1 (31.58) and ROUGE-L (27.86) scores among all approaches. This performance surpasses ExMore and ExMoreOCR by several points in n-gram overlap measures. However, in terms of SentBERT similarity, which better reflects embedding-level alignment, ExMoreOCR attains the highest score (59.57), indicating that the tri-modal design contributes to stronger sentence-level semantic correspondence with reference explanations.

These findings suggest that recent general purpose VLMs can rival or exceed specialized multimodal architectures in surface-level textual similarity, even without fine-tuning. Nonetheless, methods like ExMoreOCR may remain advantageous in capturing nuanced semantic relationships, particularly when additional OCR-derived context is available. Overall, this comparison underscores both the progress in pretrained VLM capabilities and the continued relevance of purpose-built architectures when applied to explanation generation tasks.

5. Conclusion

This study systematically examined the capabilities of contemporary VLMs for the challenging task of multimodal sarcasm explanation. Addressing RQ1, our experiments revealed that several general purpose VLMs, particularly SmolVLM-500M and Phi3.5-Vision, achieved competitive or superior performance compared to specialized architectures, even in zero-shot configurations. This suggests that recent advances in large-scale pretraining can enable effective cross-modal understanding without extensive task-specific fine-tuning. Regarding RQ2, the analysis demonstrated that while metrics such as ROUGE, BERTScore, and SentenceBERT provide valuable quantitative insights, they may overestimate explanation quality in cases of superficial textual similarity, failing to detect contradictions or factual inaccuracies. For instance, outputs conveying the opposite meaning of the reference still achieved high similarity scores, highlighting a limitation in relying solely on automatic measures to evaluate explanatory adequacy. In response to RQ3, results showed that larger models with higher parameter counts did not consistently outperform smaller models across all metrics and settings. Although some larger models obtained marginal gains in certain measures, others were surpassed by more compact architectures, underscoring that parameter size alone is not a reliable predictor of explanation quality in this domain.

The key contributions of this work include an extensive benchmarking of diverse VLMs under zero-shot and few-shot learning for multimodal sarcasm explanation, a critical evaluation of widely used similarity metrics that illustrates both their utility and limitations, and empirical evidence that challenges the assumption that larger parameter counts consistently guarantee superior performance on explanatory tasks. These findings have several implications. For researchers, they emphasize the need to complement automatic evaluation with human judgment or contradiction-aware metrics to obtain a more accurate assessment of explanation quality. For practitioners, the results suggest that deploying smaller, more efficient VLMs may achieve comparable outcomes to larger models while reducing computational costs. Nevertheless, the study has certain limitations. The evaluation was constrained to a finite set of models and metrics, and did not incorporate human annotations to validate semantic correctness. In future work, we plan to incorporate human evaluation to complement automatic metrics, such as using Likert-scale ratings to assess the relevance and clarity of sarcasm explanations. Additionally, human annotators could help identify subtle contradictions or mismatches between the generated explanation and the multimodal context. Additionally, this study focused exclusively on English-language explanations and a single dataset. Future work could explore the performance of multimodal sarcasm explanation models in multilingual settings, which would provide insights into cross-cultural and linguistic variations in sarcasm interpretation. Several recent models such as LLaVA, Sa2va, Qwen-VL, Deepseek-VL (English/Chinese), Granite Vision, TBAC VLR-1, Phi Vision, and Gemma 3 offer multilingual or cross-lingual capabilities that could be leveraged for such analysis. Future research directions include expanding evaluation to multilingual datasets, developing improved metrics capable of detecting negation and factual inconsistencies, and exploring fine-tuning strategies to enhance explanatory performance in diverse linguistic contexts. Investigating hybrid approaches that combine pretrained VLMs with specialized modules for factual verification may also offer promising avenues to improve reliability and trustworthiness in automated explanation generation.

Funding Information

The authors declare no funding was received for this study.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Ethical Approval

This study did not involve human or animal subjects.

Data Availability

- The dataset used is available at [https://github.com/LCS2-IIITD/Multimodal-Sarcasm-Explanation-MuSE].

References

[1] P. Desai, T. Chakraborty, and M. S. Akhtar, "Nice Perfume. How Long Did You Marinate in It? Multimodal Sarcasm Explanation," *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 10, pp. 10563–10571, Jun. 2022, doi: 10.1609/aaai.v36i10.21300.

[2] P. Goel, D. S. Chauhan, and M. S. Akhtar, "Target-Augmented Shared Fusion-based Multimodal Sarcasm Explanation Generation," Feb. 11, 2025, *arXiv*: arXiv:2502.07391. doi: 10.48550/arXiv.2502.07391.

[3] B. Liang *et al.*, "Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, S. Muresan, P. Nakov, and A. Villavicencio, Eds., Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 1767–1777. doi: 10.18653/v1/2022.acl-long.124.

[4] A. Rahma, S. S. Azab, and A. Mohammed, "A Comprehensive Survey on Arabic Sarcasm Detection: Approaches, Challenges and Future Trends," *IEEE Access*, vol. 11, pp. 18261–18280, 2023, doi: 10.1109/ACCESS.2023.3247427.

[5] B. Yao, Y. Zhang, Q. Li, and J. Qin, "Is sarcasm detection a step-by-step reasoning process in large language models?," *Proc. AAAI Conf. Artif. Intell.*, vol. 39, no. 24, pp. 25651–25659, Apr. 2025.

[6] X. Wang *et al.*, “Elevating Knowledge-Enhanced Entity and Relationship Understanding for Sarcasm Detection,” *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 6, pp. 3356–3371, Jun. 2025, doi: 10.1109/TKDE.2025.3547055.

[7] M. Tomar, A. Tiwari, T. Saha, and S. Saha, “Your tone speaks louder than your face! Modality Order Infused Multi-modal Sarcasm Detection,” in *Proceedings of the 31st ACM International Conference on Multimedia*, in MM ’23. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 3926–3933. doi: 10.1145/3581783.3612528.

[8] A. Bhat and A. Chauhan, “A Deep Learning based approach for MultiModal Sarcasm Detection,” in *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Dec. 2022, pp. 2523–2528. doi: 10.1109/ICAC3N56670.2022.10074506.

[9] B. Yu, H. Wang, and Z. Xi, “Multifaceted and deep semantic alignment network for multimodal sarcasm detection,” *Knowl.-Based Syst.*, vol. 301, p. 112298, Oct. 2024, doi: 10.1016/j.knosys.2024.112298.

[10] “A Multi-View Interactive Approach for Multimodal Sarcasm Detection in Social Internet of Things with Knowledge Enhancement.” Accessed: Jun. 28, 2025. [Online]. Available: https://www.mdpi.com/2076-3417/14/5/2146

[11] W. Zhong, Z. Zhang, Q. Wu, Y. Xue, and Q. Cai, “A Semantic Enhancement Framework for Multimodal Sarcasm Detection,” *Mathematics*, vol. 12, no. 2, Art. no. 2, Jan. 2024, doi: 10.3390/math12020317.

[12] “SarcNet: A Multilingual Multimodal Sarcasm Detection Dataset - ACL Anthology.” Accessed: Jun. 28, 2025. [Online]. Available: https://aclanthology.org/2024.lrec-main.1248/

[13] S. Farabi, T. Ranasinghe, D. Kanojia, Y. Kong, and M. Zampieri, “A Survey of Multimodal Sarcasm Detection,” in *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, Aug. 2024, pp. 8020–8028. doi: 10.24963/ijcai.2024/887.

[14] J. Hu, Y. Lyu, Y. Xue, F. Li, and Q. Cai, “Representation and Granularity Joint Alignment Framework for Multimodal Sarcasm Detection on Social Media,” in *Database Systems for Advanced Applications*, M. Onizuka, J.-G. Lee, Y. Tong, C. Xiao, Y. Ishikawa, S. Amer-Yahia, H. V. Jagadish, and K. Lu, Eds., Singapore: Springer Nature, 2024, pp. 243–253. doi: 10.1007/978-981-97-5575-2_17.

[15] P. Wang *et al.*, “S^3 Agent: Unlocking the Power of VLLM for Zero-Shot Multi-modal Sarcasm Detection,” *ACM Trans. Multimed. Comput. Commun. Appl.*, p. 3690642, Aug. 2024, doi: 10.1145/3690642.

[16] B. Tang, B. Lin, H. Yan, and S. Li, “Leveraging Generative Large Language Models with Visual Instruction and Demonstration Retrieval for Multimodal Sarcasm Detection,” in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 1732–1742. doi: 10.18653/v1/2024.naacl-long.97.

[17] “[2503.18681] Commander-GPT: Fully Unleashing the Sarcasm Detection Capability of Multi-Modal Large Language Models.” Accessed: Jun. 28, 2025. [Online]. Available: https://arxiv.org/abs/2503.18681?utm_source=chatgpt.com

[18] "(PDF) IRONIC: Coherence-Aware Reasoning Chains for Multi-Modal Sarcasm Detection." Accessed: Jun. 28, 2025. [Online]. Available: https://www.researchgate.net/publication/391991393_IRONIC_Coherence-Aware_Reasoning_Chains_for_Multi-Modal_Sarcasm_Detection?utm_source=chatgpt.com

[19] D. Engin and Y. Avrithis, "Zero-Shot and Few-Shot Video Question Answering with Multi-Modal Prompts," in *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, Paris, France: IEEE, Oct. 2023, pp. 2797–2802. doi: 10.1109/ICCVW60793.2023.00298.

[20] G. Yong, K. Jeon, D. Gil, and G. Lee, "Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model," *Comput.-Aided Civ. Infrastruct. Eng.*, vol. 38, no. 11, pp. 1536–1554, 2023, doi: 10.1111/mice.12954.

[21] Y. Lan, X. Li, X. Liu, Y. Li, W. Qin, and W. Qian, "Improving Zero-shot Visual Question Answering via Large Language Models with Reasoning Question Prompts," in *Proceedings of the 31st ACM International Conference on Multimedia*, in MM '23. New York, NY, USA: Association for Computing Machinery, Oct. 2023, pp. 4389–4400. doi: 10.1145/3581783.3612389.

[22] L. Sun, L. Wang, J. Sun, and T. Okatani, "Prompt Prototype Learning Based on Ranking Instruction For Few-Shot Visual Tasks," in *2023 IEEE International Conference on Image Processing (ICIP)*, Oct. 2023, pp. 3235–3239. doi: 10.1109/ICIP49359.2023.10222039.

[23] D. Ding *et al.*, "Multi-Modal Sarcasm Detection with Prompt-Tuning," in *2022 6th Asian Conference on Artificial Intelligence Technology (ACAIT)*, Dec. 2022, pp. 1–8. doi: 10.1109/ACAIT56212.2022.10137937.

[24] Y. Liu, R. Zhang, Y. Fan, J. Guo, and X. Cheng, "Prompt Tuning with Contradictory Intentions for Sarcasm Recognition," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, A. Vlachos and I. Augenstein, Eds., Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 328–339. doi: 10.18653/v1/2023.eacl-main.25.

[25] T. An, P. Yan, J. Zuo, X. Jin, M. Liu, and J. Wang, "Enhancing Cross-Lingual Sarcasm Detection by a Prompt Learning Framework with Data Augmentation and Contrastive Learning," *Electronics*, vol. 13, no. 11, Art. no. 11, Jan. 2024, doi: 10.3390/electronics13112163.

[26] S. Jana, A. Dey, and R. S. Sanasam, "Continuous Attentive Multimodal Prompt Tuning for Few-Shot Multimodal Sarcasm Detection," in *Proceedings of the 28th Conference on Computational Natural Language Learning*, L. Barak and M. Alikhani, Eds., Miami, FL, USA: Association for Computational Linguistics, Nov. 2024, pp. 314–326. doi: 10.18653/v1/2024.conll-1.25.

[27] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 34892–34916, Dec. 2023.

[28] A. Marafioti *et al.*, "SmolVLM: Redefining small and efficient multimodal models," Apr. 07, 2025, *arXiv*: arXiv:2504.05299. doi: 10.48550/arXiv.2504.05299.

[29] P. K. A. Vasu, F. Faghri, C. L. Li, C. Koc, N. True, A. Antony, *et al.*, "FastVLM: Efficient vision encoding for vision language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 19769–19780.

[30] H. Yuan *et al.*, "Sa2VA: Marrying SAM2 with LLaVA for Dense Grounded Understanding of Images and Videos," Feb. 13, 2025, *arXiv*: arXiv:2501.04001. doi: 10.48550/arXiv.2501.04001.

[31] H. Lu *et al.*, "DeepSeek-VL: Towards Real-World Vision-Language Understanding," Mar. 11, 2024, *arXiv*: arXiv:2403.05525. doi: 10.48550/arXiv.2403.05525.

[32] S. Bai *et al.*, “Qwen2.5-VL Technical Report,” Feb. 19, 2025, *arXiv*: arXiv:2502.13923. doi: 10.48550/arXiv.2502.13923.

[33] G. V. Team *et al.*, “Granite Vision: a lightweight, open-source multimodal model for enterprise Intelligence,” Feb. 14, 2025, *arXiv*: arXiv:2502.09927. doi: 10.48550/arXiv.2502.09927.

[34] L. Wei *et al.*, “Advancing Multimodal Reasoning via Reinforcement Learning with Cold Start,” May 28, 2025, *arXiv*: arXiv:2505.22334. doi: 10.48550/arXiv.2505.22334.

[35] M. Abdin *et al.*, “Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone,” Aug. 30, 2024, *arXiv*: arXiv:2404.14219. doi: 10.48550/arXiv.2404.14219.

[36] G. Team *et al.*, “Gemma 3 Technical Report,” Mar. 25, 2025, *arXiv*: arXiv:2503.19786. doi: 10.48550/arXiv.2503.19786.

[37] C. Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. ACL Workshop Text Summarization Branches Out*, Jul. 2004, pp. 74–81.

[38] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “BERTScore: Evaluating Text Generation with BERT,” Feb. 24, 2020, *arXiv*: arXiv:1904.09675. doi: 10.48550/arXiv.1904.09675.

[39] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Aug. 27, 2019, *arXiv*: arXiv:1908.10084. doi: 10.48550/arXiv.1908.10084.