Can Large Vision Language Models Understand Sarcasm?

Xinyu Wang , Yue Zhang
The University of Texas at Dallas
cpwxyxwcp@gmail.com, yue.zhang@utdallas.edu

Abstract

Sarcasm is a complex linguistic phenomenon that involves a disparity between literal and intended meanings, making it challenging for sentiment analysis and other emotion-sensitive tasks. While traditional sarcasm detection methods primarily focus on text, recent approaches have incorporated multimodal information. However, the application of Large Visual Language Models (LVLMs) in Multimodal Sarcasm Analysis (MSA) remains underexplored. In this paper, we evaluate LVLMs in MSA tasks, specifically focusing on Multimodal Sarcasm Detection and Multimodal Sarcasm Explanation. Through comprehensive experiments, we identify key limitations, such as insufficient visual understanding and a lack of conceptual knowledge. To address these issues, we propose a training-free framework that integrates in-depth object extraction and external conceptual knowledge to improve the model's ability to interpret and explain sarcasm in multimodal contexts. The experimental results on multiple models show the effectiveness of our proposed framework.

1 Introduction

Sarcasm is a linguistic phenomenon where the intended meaning opposes the literal interpretation, commonly used to convey sharp criticism or mockery toward someone or something. Understanding sarcasm is important for tasks like sentiment analysis, social media analysis, and customer feedback service, as these tasks require accurately identifying and interpreting people's emotions. Traditional sarcasm detection methods [Joshi *et al.*2016, Amir *et al.*2016] primarily focus on text modality. With the development of multimedia, recent approaches [Bouazizi and Ohtsuki2016a, Cai *et al.*2019, Qiao *et al.*2023] shift research attention into utilizing multimodal information to conduct Multimodal Sarcasm Analysis (MSA).

Despite the existing advancements in MSA [Schifanella *et al.*2016, Desai *et al.*2022], the performance of LVLMs in MSA remains unexplored. Recently, LVLMs have been evaluated on various tasks, such as VQA [Fu *et al.*2023] and image captioning [Dong *et al.*2024], demonstrating impressive

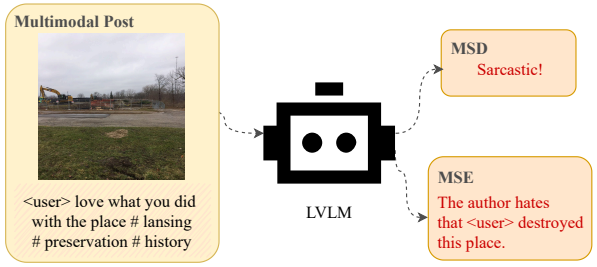

Figure 1: Multimodal Sarcasm Analysis, including MSD and MSE.

capabilities in the understanding of both visual and textual modalities [Fu *et al.*2023,Xu *et al.*2023,Bai *et al.*2023]. MSA is a non-trivial task because it involves understanding subtle cultural, emotional, and contextual nuances that are not always explicitly stated. Additionally, the task requires not only a deep understanding of both textual and visual information but also the ability to effectively leverage and integrate these modalities, further increasing the complexity of the problem. Due to the above reason, exploring the performance of LVLMs on MSA is vital for comprehensively evaluating their abilities on multimodal understanding.

Therefore, our first research question is ***RQ1: what is the zero-shot performance on MSA for LVLMs?*** To explore this research question, we assess LVLMs' performance on two key MSA tasks: Multimodal Sarcasm Detection (MSD) and Multimodal Sarcasm Explanation (MSE). Through comprehensive experiments, we found that LVLMs show poor zero-shot performance on MSA tasks. Therefore, we then explore the second research question ***RQ2: how to improve the performance on MSA for LVLMs without fine-tuning?*** Unlike other training-based methods [Ding *et al.*2022], we focus on finding a train-free method to improve the effect. We found that poor ability stems from limited visual understanding and a lack of conceptual knowledge. To address these issues, we propose an effective framework that enhances model performance by integrating in-depth object extraction and external conceptual knowledge, thereby improving sarcasm interpretation and explanation in multimodal contexts.

Our contributions can be summarized as:

- We categorize MSA into classification and explanation

tasks, with the corresponding subtasks being MSD and MSE, and evaluate the capabilities of LVLMs on these two tasks to showcase their zero-shot cross-modal analysis abilities.

- We revisit previous studies on LVLMs and analyze the limitations of LVLMs in handling sarcasm, identifying two key challenges: limited visual capabilities and insufficient conceptual knowledge.
- We propose a multi-source semantic-enhanced multimodal sarcasm understanding framework that improves LVLMs' sarcasm understanding by incorporating external knowledge sources, including detailed object and conceptual knowledge. The experimental results across multiple LVLMs show the effectiveness of our framework. Furthermore, this method provides a new perspective for extending LVLMs in complex multimodal tasks.

2 Test Task

To methodically assess the MSA abilities of LVLMs, we conduct a comprehensive evaluation focusing on their understanding and grounding capabilities in sarcasm detection and explanation. Specifically, this study addresses two tasks: Multimodal Sarcasm Detection (MSD) [Bouazizi and Ohtsuki2016b], and Multimodal Sarcasm Explanation (MSE) [Desai *et al.*2022].

2.1 Multimodal Sarcasm Detection

Task Formulation Suppose that we have a set of N testing samples $\mathcal{D} = \{s^1, s^2, \cdots, s^N\}$. Each samples $s^i = (\mathcal{T}^i, I^i, Y^i)$ involves three elements. Here, $\mathcal{T}^i$ denotes the textual sentence, I^i denotes the image, and Y^i is the ground truth label for the i-th sample. The MSD task aims to test whether a model $\mathcal{F}$ can precisely identify sarcasm in a given text and its attached image as follows,

$$\hat{Y}^i = \mathcal{F}(\mathcal{T}^i,\ I^i), \tag{1}$$

where $\hat{Y}^i$ is the binary classification prediction result of $\mathcal{F}$.

Metrics To evaluate the performance of LVLMs in the MSD task, we utilize Accuracy and F1-Score as the evaluation metrics.

2.2 Multimodal Sarcasm Explanation

Suppose we have a testing dataset $\mathcal{D}$ composed of N samples, *i.e.,* $\mathcal{D} = \{d_1, d_2, \cdots, d_N\}$. Each sample $d_i = \{T_i, I_i, Y_i\}$, where T_i denotes the input sentence, I_i is the input image, and Y_i denotes the target explanation text. The target of this task is to test whether a model $\mathcal{F}$ is able to generate the sarcasm explanation based on the given multimodal input as follows,

$$\hat{Y}_i = \mathcal{F}(T_i, I_i), \tag{2}$$

where $\hat{Y}_i$ is the generated explanation text by $\mathcal{F}$.

Metrics For evaluating the performance of LVLMs in the Multimodal Sarcasm Explanation (MSE) task, we utilize BLEU [Papineni *et al.*2002], ROUGE [Lin2004], and METEOR [Banerjee and Lavie2005], which are typically used in explanation tasks.

Table 1: Evaluation of LVLMs on the MSD task. The table shows accuracy and F1 scores for each model, with the best results highlighted in bold.

Model	Accuracy (%)	F1 (%)
LLaVA	41.1	57.4
MiniGPT	44.2	54.5
InstructBLIP	42.5	55.2
GPT-4o	**65.9**	**68.6**

Table 2: Evaluation of LVLMs on the MSE task. The table presents BLEU (B1, B2, B3, B4), ROUGE (RL, R1, R2), and Mentor scores for each model, with the highest scores in each category highlighted in bold

Model	BLEU				Rouge			Mentor
	B1	B2	B3	B4	RL	R1	R2	
LLaVA	8.285	4.689	3.120	2.203	**14.500**	14.488	5.068	**26.628**
InstructBLIP	**9.617**	**7.265**	**5.823**	**4.773**	7.738	9.133	**5.183**	7.901
MiniGPT	7.252	2.928	1.587	1.007	12.295	**16.441**	2.403	7.776
GPT-4o	5.803	3.423	2.315	1.638	10.778	12.143	4.463	25.115

3 What is Performance of LVLMs on Sarcasm Tasks?

3.1 Datasets

For the MSD and MSE tasks, we selected the testing sets of MSDD [Cai *et al.*2019] and MORE [Desai *et al.*2022] as evaluation sets, respectively. MSDD is comprised of multimodal posts from Twitter, where each post incorporates textual content and an accompanying image. Each post is assigned a label from the predefined set {sarcastic, unsarcastic}. MORE dataset collected sarcastic posts from existing multimodal sarcasm detection datasets. The researchers carefully checked the collected posts and annotated the explanation for each post. Statistics of our testing sets are summarized in Table 3.

3.2 Models

To gain a comprehensive understanding of the current state of LVLMs in sarcasm analysis, we select 3 open-source LVLMs and 1 closed-source LVLM: (1) **LLaVA-v1.5** [Liu *et al.*2023] is an enhanced version of LLaVA integrates the visual encoder CLIP with the language model LLaMA, optimized for comprehensive visual and linguistic understanding and finally refined through instruction tuning using image-based linguistic data generated by GPT-4 [OpenAI2023]. We select llava-v1.5-7b [1] for evaluation. (2) **MiniGPT** [Zhu *et al.*2024] is an

[1]https://github.com/haotian-liu/LLaVA.

Table 3: Statistics of our evaluation sets.

Task	Sources	Distribution	Total
MSD	MSDD	Sarcastic 959 Unsarcastic 1450	2409
MSE	MORE	Caption Avg.length = 19.43 Explanation Avg.length = 15.08	352

open-source LVLM that aligns a frozen visual encoder with a frozen language model LLaMA [Touvron *et al.*2023], refined through instruction tuning using some instruction datasets. We utilize minigptv2-llama-7b [2] for evaluation. (3) **InstructBLIP** [Dai *et al.*2023] is an open-source LVLM based on a pre-trained BLIP-2 model, achieving multimodal capabilities through visual-language instruction adjustment. We utilize the InstructBLIP-vicuna-7b [3] for testing. (4) **GPT-4o** [OpenAI2023] is a version of the GPT-4 model designed for optimized performance in both language and vision tasks. It has been refined through stages of pre-training, instruction tuning, and reinforcement learning from human feedback. We utilize gpt-4o-mini version for evaluation.

3.3 Result

In Table 1 and 2, we showcase the performance of all 4 LVLMs tested in the zero-shot setting on MSE and MSD tasks. Based on the results in Table 1 and 2, we draw the following observations: On the whole, GPT-4o shows the strongest performance on the MSD task, achieving the highest accuracy and F1 score, while it demonstrated poor performance in sarcasm explanation generation. Other models perform poorly on the sarcasm detection task and show different capabilities on the sarcasm explanation task, highlighting the need for further improvements to help them achieve better results.

4 Multi-source Semantic enhanced Multimodal Sarcasm Understanding

4.1 Overview

To enhance performance on multimodal sarcasm detection and explanation tasks, we revisited previous studies on LVLMs and identified two primary challenges that may have contributed to the poor performance of these LVLMs: 1) **Lack of Vision Ability**: LVLMs may produce hallucinations [Li *et al.*2023, Jing *et al.*2023a, Lu *et al.*2023, Zhou *et al.*2024a], especially in fine-grained objects, which leads to their limited capabilities in visual understanding. 2) **Lack of Potential External Knowledge**: Due to the limited nature of the training data, LVLMs may not be able to make connections between existing visual entities and related concepts [Liu *et al.*2024, Xuan *et al.*2024, Wang *et al.*2024], such as sentiment knowledge, which is important for sarcasm understanding.

To address both challenges, we propose a multi-source semantic enhanced multimodal sarcasm understanding framework, which mainly consists of three steps: fine-grained object extraction, external knowledge acquisition, and result generation. Fig. 2 illustrates the overview of our method. In this framework, we focus on enhancing the performance of LVLMs in MSD and MSE tasks by introducing recognized objects with attributes and external knowledge. This approach aims at improving the models' ability to understand and interpret the potential meaning conveyed in images and text, especially in cases where the context is complex or subtle.

[2]https://github.com/Vision-CAIR/MiniGPT-4.

[3]https://github.com/salesforce/LAVIS/tree/main/projects/instructblip.

4.2 Method

Fine-grained Object Extraction We extract objects and their attributes from the image and incorporate them into the prompt, providing LVLMs with a more comprehensive context to leverage for sarcasm understanding. Specifically, we utilize Fast-RCNN [Girshick2015] as our chosen fine-grained visual object recognizer. Fast-RCNN is well-regarded for its ability to accurately detect and describe objects within an image. It can identify and extract a set of objects with key descriptive attributes, such as shape, color, and other visual features, that contribute to a comprehensive understanding of the image's content. By integrating these extracted objects with their attributes into the prompt, we can provide the LVLMs with more detailed information, thereby improving their capacity to analyze and interpret the potential meaning conveyed in the image.

External knowledge Acquisition However, simply extracting objects and their attributes may not always result in accurate sarcasm analysis. This is because LVLMs might struggle to fully grasp the underlying properties or the contextual relevance of the extracted objects and input text. To mitigate this issue, we propose to introduce external knowledge derived from the attributes of the objects and text. The idea is to bridge the gap between the visual features of the objects and the deeper, more abstract concepts they represent. To achieve this, we employ ConceptNet [Speer *et al.*2017] as our Knowledge Generator. ConceptNet is a powerful tool for capturing and organizing human knowledge, particularly in the field of concepts and their interrelationships. By feeding the attributes of the extracted objects and input text into ConceptNet, we can identify the most relevant concepts associated with them, *i.e.,* their neighboring concepts. These concepts represent the underlying ideas or meanings that the objects and input text might convey in a given context.

Result Generation After acquiring these concepts, we incorporate them into the prompt to provide the LVLMs with a more nuanced understanding of the objects and input text. This enables the models to interpret the sarcastic meaning expressed in both the image and the text more accurately. We use the enhanced prompt for the final sarcasm analysis, incorporating detailed object information and relevant external concepts.

4.3 Result

We compare our method with existing baselines on each task and report the results in Table 5 and Table 4. After introducing fine-grained objects and additional conceptual knowledge, our method has achieved the best results on both tasks. Impressively, in the MSD task, on GPT-4o, we achieved 75.3% accuracy, a 14.2% relative improvement over the baseline. In particular, on the MSE task, our method achieved great improvement on InstructBLIP and MiniGPT. In addition, we found that LLaVA, which had a poor baseline performance, was improved on both tasks after applying our

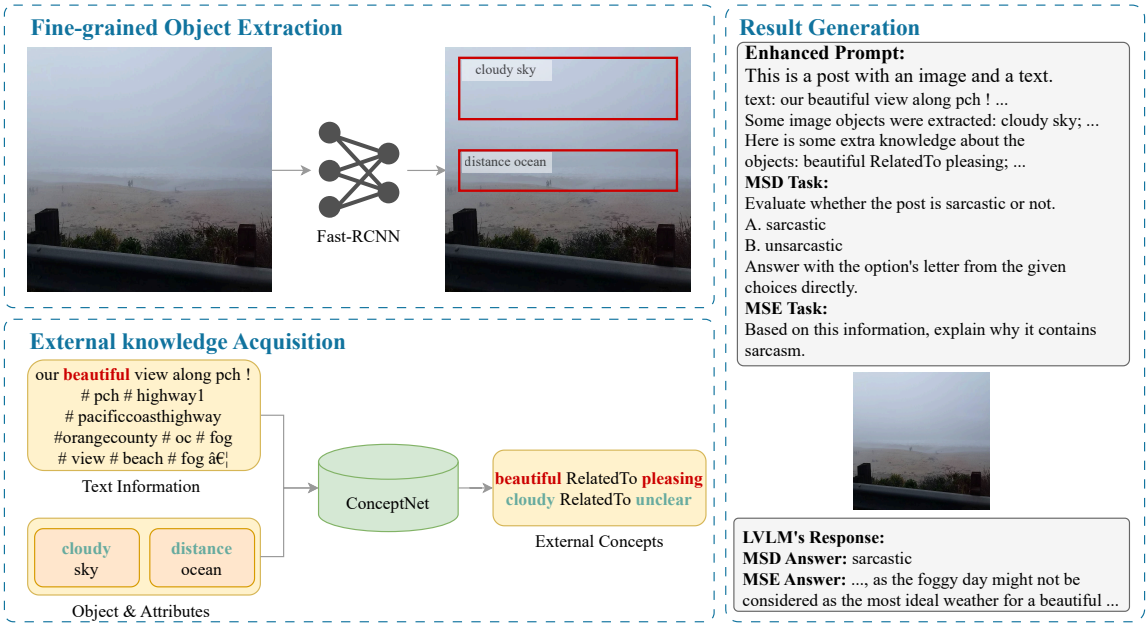

Figure 2: **The multi-source semantic enhanced sarcasm understanding framework**: 1) **Fine-grained object extraction** using Fast-RCNN to detect objects from images, 2) **External knowledge acquisition**, linking text features and image objects to external concepts through ConceptNet, enriching understanding by associating attributes. 3) **Result generation** for sarcasm detection and explanation. The framework leverages these components to enhance sarcasm comprehension in multimodal contexts.

method, which shows that our method effectively supplements visual capabilities and provides sufficient extra knowledge for the LVLMs to assist them in completing the sarcasm analysis task.

5 Related Work

5.1 Evaluation in LVLMs.

LVLMs emerged as a focal point of interest due to their remarkable capabilities in handling diverse multimodal tasks. Researchers have developed various LVLMs [Liu *et al.*2023, Zhu *et al.*2024, Dai *et al.*2023], which generally consist of an image encoder and a text decoder derived from pre-trained models and a text-image alignment module. These LVLMs exhibit excellent generalization abilities and can be applied to many scenarios. As multimodal research deepens, some studies began focusing on evaluating the zero-shot capabilities of LVLMs [Lu *et al.*2023], forming a series of benchmarks [Yang *et al.*2024, Lin *et al.*2024]. In addition, more and more studies have found that the existing LVLMs have some defects, such as Hallucinations [Jing *et al.*2023a], and Overfitting [Xu *et al.*2023]. To mitigate these issues, techniques like Chain-of-Thought [Zhang *et al.*2024, Zheng *et al.*2023] and In-Context-Learning [Zhou *et al.*2024b, Shukor *et al.*2024] are increasingly employed to enhance model performance during inference. Despite extensive research into the evaluation and enhancement of LVLMs, there is a limited focus on assessing and improving their performance for sarcasm analysis tasks.

5.2 Multimodal Sarcasm Detection and Explanation.

In the sarcasm analysis task, there are two key sub-tasks, including sarcasm detection and sarcasm explanation. Recent studies [Desai *et al.*2022, Cai *et al.*2019, Tang *et al.*2024, Chen *et al.*2024, Wen *et al.*2023, Qiao *et al.*2023, Yue *et al.*2023, Wu *et al.*2021, Jing *et al.*2023b] have concentrated on developing specialized training methods and models to achieve state-of-the-art results in these tasks. Notably, some research efforts, such as those presented in [Jing *et al.*2023b, Wang *et al.*2024], have demonstrated that incorporating additional external knowledge—such as contextual information or background knowledge—can significantly enhance model performance by providing a richer understanding of the sentiment conveyed. Different from the existing works, we focus on evaluating and improving sarcasm understanding ability in zero-shot scenarios.

6 Conclusion

In this paper, we evaluated the zero-shot capabilities of Large Vision-Language Models (LVLMs) on the core tasks of Multimodal Sarcasm Analysis (MSA), specifically focusing on sarcasm detection and explanation. Our findings reveal that LVLMs perform poorly in understanding multimodal sarcasm content, particularly due to limitations in visual semantics and conceptual knowledge. To address these gaps, we proposed incorporating in-depth object information and external conceptual knowledge sources to enhance the models' performance. This approach offers a promising direction for improving the ability of LVLMs to handle complex sarcas-

Table 4: Performance comparison of LVLMs on the MSE task. The best results are highlighted in bold. * indicates that the p-value of the significance test comparing our result with the best baseline result is less than 0.01.

Model	Method	BLEU				Rouge			Mentor
		B1	B2	B3	B4	RL	R1	R2	
LLaVA	Baseline	8.285	4.689	3.120	2.203	14.500	14.488	5.068	26.628
	+Ours	**9.430***	**5.681***	**3.925***	**2.852***	**16.301***	**16.549***	**6.447***	**29.640***
InstructBLIP	Baseline	9.617	7.265	5.823	4.773	7.738	9.133	5.183	7.901
	+Ours	**14.196***	**9.198***	**6.866***	**5.343***	**22.473***	**23.386***	**11.140***	**30.380***
MiniGPT	Baseline	7.252	2.928	1.587	1.007	12.295	16.441	2.403	7.776
	+Ours	**21.173***	**11.400***	**7.276***	**4.846***	**19.000***	**23.074***	**6.202***	**16.600***
GPT-4o	Baseline	5.803	3.423	2.315	1.638	10.778	12.143	4.463	25.115
	+Ours	**8.863***	**4.989***	**3.202***	**2.148***	**14.242***	**15.475***	**5.595***	**28.717***

Table 5: Performance comparison of LVLMs on the MSD task. The table shows accuracy and F1 scores for each model using the baseline method and our method. The highest values in each category are highlighted in bold. The relative increase percentages are noted in parentheses.

Model	Method	Accuracy (%)	F1 (%)
LLaVA	Baseline	41.1	57.4
	+Ours	**60.3** (+46.8%)	**59.4** (+3.5%)
MiniGPT	Baseline	44.2	54.5
	+Ours	**50.0** (+13.2%)	**56.5** (+3.7%)
InstructBLIP	Baseline	42.5	55.2
	+Ours	**51.5** (+21.2%)	**57.9** (+4.9%)
GPT-4o	Baseline	65.9	68.6
	+Ours	**75.3** (+14.2%)	**72.1** (+5.1%)

tic content in multimodal scenarios.

References

[Amir *et al.*, 2016] Silvio Amir, Byron C. Wallace, Hao Lyu, and Paula Carvalho Mário J. Silva. Modelling context with user embeddings for sarcasm detection in social media, 2016.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL*, pages 65–72. Association for Computational Linguistics, 2005.

[Bouazizi and Ohtsuki, 2016a] Mondher Bouazizi and Tomoaki Ohtsuki. A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488, 2016.

[Bouazizi and Ohtsuki, 2016b] Mondher Bouazizi and Tomoaki Ohtsuki. A pattern-based approach for sarcasm detection on twitter. *IEEE Access*, 4:5477–5488, 2016.

[Cai *et al.*, 2019] Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *ACL*, pages 2506–2515. Association for Computational Linguistics, 2019.

[Chen *et al.*, 2024] Zixin Chen, Hongzhan Lin, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen. Cofipara: A coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models. In *ACL*, pages 9663–9687. Association for Computational Linguistics, 2024.

[Dai *et al.*, 2023] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*, 2023.

[Desai *et al.*, 2022] Poorav Desai, Tanmoy Chakraborty, and Md. Shad Akhtar. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *AAAI*, pages 10563–10571. AAAI Press, 2022.

[Ding *et al.*, 2022] Daijun Ding, Hu Huang, Bowen Zhang, Cheng Peng, Yangyang Li, Xianghua Fu, and Liwen Jing. Multi-modal sarcasm detection with prompt-tuning. 2022.

[Dong *et al.*, 2024] Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *CoRR*, abs/2405.19092, 2024.

[Fu *et al.*, 2023] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023.

[Girshick, 2015] Ross B. Girshick. Fast R-CNN. In *ICCV*, pages 1440–1448. IEEE Computer Society, 2015.

[Jing *et al.*, 2023a] Liqiang Jing, Ruosen Li, Yunmo Chen, Mengzhao Jia, and Xinya Du. FAITHSCORE: evaluating hallucinations in large vision-language models. *CoRR*, abs/2311.01477, 2023.

[Jing *et al.*, 2023b] Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. Multi-source semantic graph-based multimodal sarcasm explanation generation. In *ACL*, pages 11349–11361. Association for Computational Linguistics, 2023.

[Joshi *et al.*, 2016] Aditya Joshi, Vaibhav Tripathi, Kevin Patel, Pushpak Bhattacharyya, and Mark Carman. Are word embedding-based features useful for sarcasm detection? In *EMNLP*, pages 1006–1011. Association for Computational Linguistics, November 2016.

[Li *et al.*, 2023] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305. Association for Computational Linguistics, 2023.

[Lin *et al.*, 2024] Hongzhan Lin, Ziyang Luo, Bo Wang, Ruichao Yang, and Jing Ma. Goat-bench: Safety insights to large multimodal models through meme-based social abuse. *CoRR*, abs/2401.01523, 2024.

[Lin, 2004] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics, July 2004.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

[Liu *et al.*, 2024] Xuannan Liu, Peipei Li, Huaibo Huang, Zekun Li, Xing Cui, Jiahao Liang, Lixiong Qin, Weihong Deng, and Zhaofeng He. Fakenewsgpt4: Advancing multimodal fake news detection through knowledge-augmented lvlms. *CoRR*, abs/2403.01988, 2024.

[Lu *et al.*, 2023] Jiaying Lu, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Baochen Sun, Carl Yang, and Jie Yang. Evaluation and enhancement of semantic grounding in large vision-language models, 2023.

[OpenAI, 2023] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. ACL, 2002.

[Qiao *et al.*, 2023] Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *AAAI*, pages 9507–9515. AAAI Press, 2023.

[Schifanella *et al.*, 2016] Rossano Schifanella, Paloma de Juan, Joel R. Tetreault, and Liangliang Cao. Detecting sarcasm in multimodal social platforms. In *MM 2016*, pages 1136–1145. ACM, 2016.

[Shukor *et al.*, 2024] Mustafa Shukor, Alexandre Ramé, Corentin Dancette, and Matthieu Cord. Beyond task performance: evaluating and reducing the flaws of large multimodal models with in-context-learning. In *ICLR 2024*. OpenReview.net, 2024.

[Speer *et al.*, 2017] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451. AAAI Press, 2017.

[Tang *et al.*, 2024] Binghao Tang, Boda Lin, Haolong Yan, and Si Li. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In *NAACL*, pages 1732–1742. Association for Computational Linguistics, 2024.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971, 2023.

[Wang *et al.*, 2024] Wenbin Wang, Liang Ding, Li Shen, Yong Luo, Han Hu, and Dacheng Tao. Wisdom: Improving multimodal sentiment analysis by fusing contextual world knowledge. *CoRR*, abs/2401.06659, 2024.

[Wen *et al.*, 2023] Changsong Wen, Guoli Jia, and Jufeng Yang. Dip: Dual incongruity perceiving network for sarcasm detection. In *CVPR*, pages 2540–2550, June 2023.

[Wu *et al.*, 2021] Yang Wu, Yanyan Zhao, Xin Lu, Bing Qin, Yin Wu, Jian Sheng, and Jinlong Li. Modeling incongruity between modalities for multimodal sarcasm detection. *IEEE Multim.*, 28(2):86–95, 2021.

[Xu *et al.*, 2023] Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *CoRR*, abs/2306.09265, 2023.

[Xuan *et al.*, 2024] Keyang Xuan, Li Yi, Fan Yang, Ruochen Wu, Yi R. Fung, and Heng Ji. LEMMA: towards lvlm-enhanced multimodal misinformation detection with external knowledge augmentation. *CoRR*, abs/2402.11943, 2024.

[Yang *et al.*, 2024] Xiaocui Yang, Wenfang Wu, Shi Feng, Ming Wang, Daling Wang, Yang Li, Qi Sun, Yifei Zhang, Xiaoming Fu, and Soujanya Poria. Mm-instructeval: Zero-shot evaluation of (multimodal) large language models on multimodal reasoning tasks. *CoRR*, abs/2405.07229, 2024.

[Yue *et al.*, 2023] Tan Yue, Rui Mao, Heng Wang, Zonghai Hu, and Erik Cambria. Knowlenet: Knowledge fusion network for multimodal sarcasm detection. *Inf. Fusion*, 100:101921, 2023.

[Zhang *et al.*, 2024] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Trans. Mach. Learn. Res.*, 2024, 2024.

[Zheng *et al.*, 2023] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *NeurIPS 2023*, 2023.

[Zhou *et al.*, 2024a] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *ICLR 2024*. OpenReview.net, 2024.

[Zhou *et al.*, 2024b] Yucheng Zhou, Xiang Li, Qianning Wang, and Jianbing Shen. Visual in-context learning for large vision-language models. In *Findings of the Association for Computational Linguistics, ACL 2024*, pages 15890–15902. Association for Computational Linguistics, 2024.

[Zhu *et al.*, 2024] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *ICLR*. OpenReview.net, 2024.