S^3-MSD: Large Vision-Language Model for Explainable and Generalizable Multimodal Sarcasm Detection

Zhihong Zhu[1], Fan Zhang[2], Yunyan Zhang[1], Jinghan Sun[1], Guimin Hu[3], Hao Wu[1], Yuyan Chen[4], Bowen Xing[5], Xian Wu[1]*

[1]Tencent Jarvis Lab
[2]The Chinese University of Hong Kong
[3]University of Copenhagen
[4]Cornell University
[5]University of Science and Technology Beijing
{profzhu, kevinxwu}@tencent.com

Abstract

Multimodal sarcasm detection (MSD) aims to identify sarcasm polarity through diverse modalities (*i.e.*, image-text pairs), which gains increasing attention. While significant progress has been made, the existing approaches still face two major issues: *lack of explainability* and *weak generalizability*. In this paper, we introduce a new large vision-language model (LVLM) dubbed S^3-MSD for explainable and generalizable MSD through three key components. For explainability, we develop (1) a self-training paradigm that automatically bootstraps answers with explanations, and (2) a self-calibrating mechanism that rectifies flawed explanations. For generalizability, we design (3) a self-focusing module that amplifies visual semantic entities through preference optimization, thereby mitigating textual over-reliance. Experimental results on both in-distribution and out-of-distribution (OOD) benchmarks demonstrate that S^3-MSD consistently outperforms state-of-the-art methods in detection performance. Furthermore, the proposed S^3-MSD provides persuasive explanations, as validated by quantitative and human evaluations.

1 Introduction

Sarcasm is a sophisticated form of figurative language where the expressed sentiment contradicts the literal meaning (Gibbs 1986; Joshi, Bhattacharyya, and Carman 2017). With the explosive growth of social media, Multimodal Sarcasm Detection (MSD) has emerged as a crucial research area (Farabi et al. 2024; Ma et al. 2024a; Liang et al. 2024a), due to its importance in diverse applications such as product review analysis (Wen, Jia, and Yang 2023). Distinct from conventional text-only analysis (Li et al. 2021; Lou et al. 2021), MSD presents a formidable challenge (Qin et al. 2023): it requires discerning subtle semantic inconsistencies across different modalities (*e.g.* text and image).

To address this challenge, a group of MSD models has been proposed (Zhang et al. 2025a,b). Mainstream efforts follow a *classification-based* paradigm, leveraging backbones such as BERT (Devlin et al. 2019) for text and ViT (Dosovitskiy et al. 2020) for vision, to learn a unified latent representation for sarcasm detection. Within this paradigm, a variety of fusion strategies have been explored to capture cross-modal incongruity, including attention-based methods (Pan et al. 2020; Xu, Zeng, and Mao 2020), graph-based frameworks (Liang et al. 2022; Wei et al. 2024), knowledge-enhanced techniques (Liu, Wang, and Li 2022; Wei et al. 2025), and information-decoupling strategies (Tian et al. 2023; Qin et al. 2023; Goel, Chauhan, and Akhtar 2025). Another stream of methods has made early efforts to introduce *generation-based* LVLMs into MSD, due to their impressive performance across various multimodal applications (Liang et al. 2024b). Among these, Tang et al. (2024) proposed a CLIP-based (Radford et al. 2021) demonstration retrieval module to boost cross-modal in-context learning, obtaining promising results.

Despite substantial progress achieved, we discover that current MSD methods still suffer from two main issues:

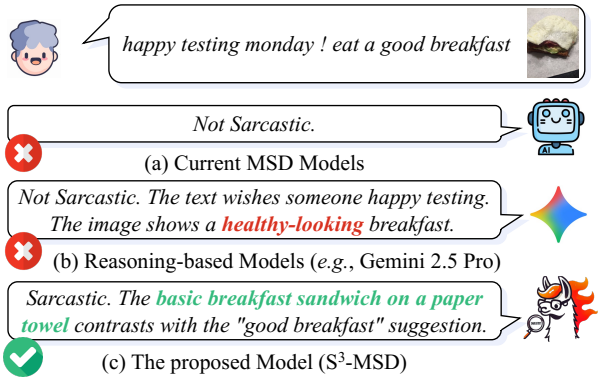


Figure 1: Comparison on a sarcastic sample where the text suggests a "*good breakfast*" while the image shows a basic sandwich on a paper towel. (a) Existing MSD models fail to capture this nuance without explanation. (b) Reasoning-based models like Gemini 2.5 Pro provide a flawed explanation by focusing only on the food's healthy appearance while overlooking the casual, understated presentation. (c) In contrast, the proposed S^3-MSD correctly identifies the sarcasm by grounding its explanation in these key visual details.

*Corresponding author.

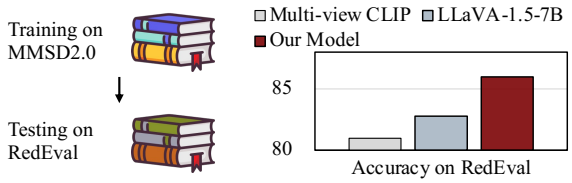


Figure 2: Performance comparison between state-of-the-art models (Qin et al. 2023; Touvron et al. 2023) and the proposed model on OOD dataset RedEval (Tang et al. 2024).

(1) *Lack of explainability*. The development for explainable MSD remains largely unexplored, yet this capability is critical for convincing and effective sarcasm detection (Jing et al. 2023). While LVLMs have the potential to generate explanations, directly applying them to MSD is non-trivial. Without task-specific supervision, even reasoning-based LVLMs also fail to capture the subtle cross-modal inconsistencies as shown in Figure 1. Given that manual annotation is prohibitively expensive and difficult to scale, *how can we generate the supervisory signals from classification-only data to teach an LVLM to reason reliably for MSD?*

(2) *Weak generalizability*. Preliminary experiments on the OOD dataset (Tang et al. 2024) with state-of-the-art *classification-based* and *generation-based* methods in Figure 2 reveal their suboptimal performance (Qin et al. 2023). This can be attributed to the fact that current MSD models rely more heavily on textual modality than on visual modality (Pan et al. 2020). As such, unreliable textual cues (*e.g.*, biased words that frequently appear in either sarcastic or non-sarcastic samples) can mislead the models, hindering their generalizability when the text distribution shifts. This raises our second research question: *How can we mitigate over-reliance on text to achieve generalizable MSD*?

To answer the above two questions, we introduce S^3-MSD, a new LVLM for explainable and generalizable MSD. For explainability, we (1) first adopt a *self-training paradigm* that leverages the pre-existing but weak reasoning ability from the LVLM to bootstrap both explanation and answer. In this manner, S^3-MSD enables the collection of both positive and negative explanations based on the golden label. Inspired by the human ability to learn from mistakes (Zhou et al. 2024; Zhang et al. 2024), we then (2) introduce a *self-calibrating mechanism* to refine flawed explanations, allowing the model to learn from erroneous reasoning.

For generalizability, we further (3) propose a *self-focusing module*, which mitigates excessive textual dependency bias by emphasizing semantic entities in the images. Concretely, we perform preference optimization (Rafailov et al. 2023) only on images, where the rejected image is derived from the original by introducing noise to visual semantic entities.

Overall, our contributions can be summarized as follows:

- We present a new LVLM-based model dubbed S^3-MSD for MSD and make the first attempt to bridge detection and explanation for multimodal sarcasm understanding.
- We introduce three core components in S^3-MSD: *self-training*, *self-calibrating*, and *self-focusing*, to address issues of lack of explainability and weak generalizability.
- Extensive experiments on both in-distribution and out-of-distribution benchmarks including quantitative comparison, qualitative analysis, human evaluation, cross-architecture, and cross-task scenarios verify the explainability and generalizability of the proposed S^3-MSD.

2 S^3-MSD

Task Formulation. Given an input image-text pair (x^v, x^t), we utilize an LVLM $\mathcal{M}(\cdot)$ to perform two tasks simultaneously: (1) sarcasm detection, where it predicts a label $a \in \{0, 1\}$, with $a = 1$ indicating sarcasm and $a = 0$ indicating non-sarcasm; (2) explanation generation, where it produces an explanation e to justify its prediction. Formally, S^3-MSD operates as follows: $\mathcal{M}(x^v, x^t) \rightarrow (e, a)$.

In the following, we detail the proposed S^3-MSD, whose overall framework is illustrated in Figure 3.

2.1 Self-training

Motivation. Given the challenge of obtaining high-quality explanation data, directly fine-tuning the LVLM for explainable MSD remains impractical. Inspired by Zelikman et al. (2022), we propose leveraging the LVLM's inherent yet limited reasoning ability to iteratively augment pairs from the original available MSD dataset $\mathcal{D}$, thereby enabling the model to self-train. At each iteration k, the model $\mathcal{M}(\cdot)$ first processes each input image-text pair $(x_i^v, x_i^t) \in \mathcal{D}_{i=1}^{|D|}$ to generate both prediction and explanation simultaneously:

$$y_i = (e_i^{(k)}, a_i^{(k)}) = \mathcal{M}_k(x_i^v, x_i^t), \tag{1}$$

where $a_i^{(k)}$ is the sarcasm prediction extracted from model response y_i, and $e_i^{(k)}$ represents the supported explanation.

Given that correct predictions often correlate with higher-quality explanations, we retain samples with correct answers to form the augmented set for self-training $\mathcal{D}_{st}^{(k)}$:

$$\mathcal{D}_{st}^{(k)} = \left\{ (x_i^v, x_i^t, e_i^{(k)}, a_i^{(k)}) \mid a_i^{(k)} = \hat{a}_i \right\}_{i=1}^{|D|}, \tag{2}$$

where $\hat{a}_i$ denotes the ground-truth label of i-th sample.

Subsequently, we fine-tune the model $\mathcal{M}_k(\cdot)$ on the newly augmented dataset $\mathcal{D}_{st}^{(k)}$ using supervised fine-tuning (SFT) with a standard negative log-likelihood objective:

$$\mathcal{L}_{sft} = -\sum_{(x_i^v, x_i^t, y_i) \in \mathcal{D}_{st}^{(k)}} \sum_j \log p_{\mathcal{M}_k}(y_{i,j} \mid y_{i,<j}, x_i^v, x_i^t), \tag{3}$$

where j indexes the tokens in the output sequence y_i. This iterative self-training process continues, generating a new response for each sample using the newly fine-tuned model, progressively improving sarcasm detection and explanation quality until performance converges.

2.2 Self-calibrating

Motivation. While *self-training* enhances the model by leveraging positive samples with accurately predicted explanations, it overlooks negative ones. Incorrect predictions

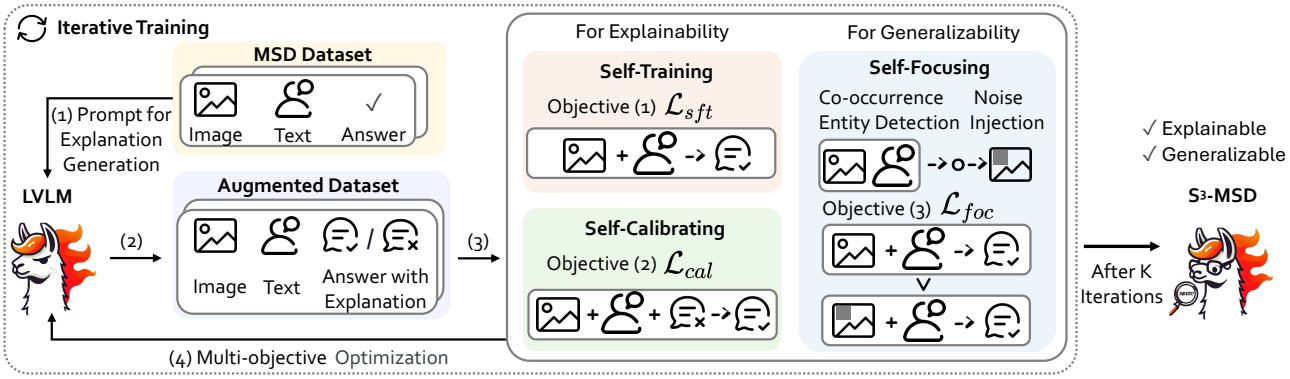


Figure 3: Overview of the proposed S^3-MSD framework. At each iteration k, we leverage the LVLM to bootstrap an augmented dataset from the original MSD dataset. The LVLM then undergoes *self-training* (§2.1), *self-calibration* (§2.2), and *self-focusing* (§2.3) optimization, to achieve explainable and generalizable S^3-MSD.

constitute most sampled instances and likewise offer valuable insights for model improvement. Though direct preference optimization (DPO) (Rafailov et al. 2023) is a potential approach for utilizing textual positive-negative pairs, our controlled experiments reveal its suboptimal performance in handling multimodal incongruity (*cf.* §3). Building upon human error-driven learning mechanisms (Zhou et al. 2024), we propose a *self-calibration mechanism* that systematically transforms prediction errors into explanatory signals.

Specifically, for each image-text pair $x_i^{v/t}$ at the k-th iteration, we identify samples where the model has produced both correct and incorrect sarcasm predictions at different iterations up to k, forming a self-calibration set $\mathcal{D}_{sc}^{(k)}$ as:

$$\mathcal{D}_{sc}^{(k)} = \Big\{(x_i^v, x_i^t, y_i^+, y_i^-) \,\Big|\, \exists\, m, n \le k : a_i^{(m)} = \hat{a}_i \wedge a_i^{(n)} \neq \hat{a}_i \Big\}_{i=1}^{|D|}, \tag{4}$$

where y_i^+ and y_i^- denote the generated explanation-prediction pairs for correct and incorrect classifications.

To effectively incorporate feedback into the training process, we introduce a self-calibrating loss function that compels S^3-MSD to calibrate its responses based on its own mistakes. Formally, this objective $\mathcal{L}_{cal}$ is defined as:

$$\mathcal{L}_{cal} = -\sum_{(x_i^v, x_i^t, y_i^+, y_i^-) \in \mathcal{D}_{sc}^{(k)}} \sum_j \log p_{\mathcal{M}_k}(y_{i,j}^+ \mid y_{i,<j}^+, x_i^v, x_i^t, y_i^-), \tag{5}$$

where the model is encouraged to generate a more reliable response y_i^+ by conditioning on the incorrect one y_i^-. In this fashion, the model can iteratively refine its reasoning over successive iterations, mitigating erroneous interpretations.

2.3 Self-focusing

Motivation. Due to over-reliance on the textual modality, current MSD models shows weak generalizability in OOD scenarios. To this end, we shift towards the visual modality, leveraging localized emphasis to effectively capture sarcastic cues. We propose a *self-focusing module* that compels the model to attend to sarcasm-related visual entities.

We resort to DPO, a popular method for aligning language models with human preferences. Instead of learning a separate reward model, DPO directly optimizes the policy using preference data. Given a dataset of preference pairs $\{x, y^+, y^-\} \in \mathcal{D}$, where y^+ and y^- denote the preferred and dispreferred responses for input x, respectively, the probability of preferring y^+ over y^- is modeled as: $p(y^+ \succ y^-) = \sigma\big(r(x, y^+) - r(x, y^-)\big)$, where $\sigma(\cdot)$ denotes the sigmoid function. To directly optimize for this preference structure, DPO expresses the reward function in terms of the policy π_θ and a fixed reference model π_{ref} as follows: $r(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log Z(x)$, where β is a scaling factor and $Z(x)$ is the partition function. Substituting this reward formulation into the preference objective, the optimization objective of DPO can be derived as:

$$\mathcal{L} = -\mathbb{E}_{(x,y^+,y^-)\sim\mathcal{D}}\Big[\log\sigma\big(\beta \log \frac{\pi_\theta(y^+|x)}{\pi_{\text{ref}}(y^+|x)} - \beta \log \frac{\pi_\theta(y^-|x)}{\pi_{\text{ref}}(y^-|x)}\big)\Big]. \tag{6}$$

During training, the reference π_{ref} remains fixed, while the policy π_θ is updated to maximize preference alignment.

Unlike standard DPO that aligns textual responses, we reformulate the input as multimodal pairs (x^t, x^v). Let $\mathcal{D}^{(k)} = \{(x_i^t, x_i^{(v,+)}, x_i^{(v,-)}, y_i^+)\}_{i=1}^{|D^{(k)}|}$ be our constructed preference set, where (1) $x^{(v,+)}$ denotes the original image with sarcasm-relevant entities, and (2) $x^{(v,-)}$ denotes perturbed image where key entities are corrupted. Following Eq. (6), the *self-focusing* loss function can be rewritten as follows:

$$\mathcal{L}_{foc} = -\mathbb{E}_{(x_i^t, x_i^{(v,+)}, x_i^{(v,-)}, y_i^+)\sim\mathcal{D}^{(k)}}\Bigg[\log\sigma\Big(\beta \log \frac{\pi_\theta(y_i^+|x_i^{(v,+)}, x_i^t)}{\pi_{\text{ref}}(y_i^+|x_i^{(v,+)}, x_i^t)} - \beta \log \frac{\pi_\theta(y_i^+|x_i^{(v,-)}, x_i^t)}{\pi_{\text{ref}}(y_i^+|x_i^{(v,-)}, x_i^t)}\Big)\Bigg], \quad \text{subject to} \quad \exists\, m, n \le k : a_i^{(m)} = \hat{a}_i \wedge a_i^{(n)} \neq \hat{a}_i. \tag{7}$$

Here, we focus on samples that have been both previously mispredicted and correctly predicted, as they are ambiguous.

The challenge then lies in constructing the dispreferred image input $x_i^{(v,-)}$. We extract textual entities $\mathcal{E}_t$ and visual entities $\mathcal{E}_v$ using a syntax parser[1] and an object detector.[2] After that, we randomly select an entity from $\mathcal{E}_t \cap \mathcal{E}_v$; if $\mathcal{E}_t \cap \mathcal{E}_v = \varnothing$, we instead randomly select one from $\mathcal{E}_v$. We then introduce controlled noise into the bounding box of the selected entity region $x^v_{(i,r)} + \mathcal{N}(0, \sigma^2)$, where $\mathcal{N}(0, \sigma^2)$ denotes Gaussian noise with perturbation intensity controlled by σ. Finally, the perturbed region is reintegrated into the image x_i^v to form the dispreferred image input $x_i^{(v,-)}$.

Theoretical Analysis. *Let $X = (x^v, x^t)$ be input image-text pair. The model's prediction is denoted by $p(y \mid X) = f(\phi_v(x^v), \phi_t(x^t))$, where ϕ_v and ϕ_t encode the visual and textual semantics, respectively. Under a distribution shift, over-reliance on textual features ϕ_t can lead to suboptimal performance. We address this by encouraging the model to be sensitive to perturbations in key visual entities:*

$$\mathbb{E}_{x^{(v,+)} \sim \mathcal{D},\, x^{(v,-)} = x^{(v,+)} + \Delta_v} \left[\log \frac{\pi_\theta(y^+ \mid x^t, x^{(v,+)})}{\pi_\theta(y^+ \mid x^t, x^{(v,-)})} \right] \geq \delta, \tag{8}$$

where Δ_v perturbs key visual entities in x^v. This principle is implemented via our self-focusing loss $\mathcal{L}_{foc}$ (from Eq. 7), which we formulate as a constrained optimization problem:

$$\min_{\pi_\theta} \mathcal{L}_{foc} \triangleq -\mathbb{E}\left[\log \sigma\left(\beta\, \Delta \log \pi_\theta\right)\right] \quad \text{s.t. } I(y; \phi_v \mid \phi_t) \uparrow, \tag{9}$$

where $I(y; \phi_v \mid \phi_t)$ measures the information gain from visual features ϕ_v given textual features ϕ_t, and $\Delta \log \pi_\theta = \log \pi_\theta(y^+ \mid x^{(v,+)}, x^t) - \log \pi_\theta(y^+ \mid x^{(v,-)}, x^t)$.

By solving this objective, the gradient updates with respect to the textual features ϕ_t can be shown to satisfy:

$$\begin{aligned} \frac{\partial \mathcal{L}_{foc}}{\partial \phi_t} \;\propto\; & \nabla_{\phi_t} I(y; \phi_v \mid \phi_t) \\ & - \lambda \frac{\partial}{\partial \phi_t} \underbrace{\mathrm{KL}\big(p(y \mid \phi_v, \phi_t) \,\|\, p(y \mid \phi_t)\big)}_{\text{Textual Bias}}, \end{aligned} \tag{10}$$

thus suppressing spurious dependence on ϕ_t by penalizing the textual bias term, which in turn preserves reasoning grounded in visual features ϕ_v (assuming $\nabla_{\phi_t} I(\cdot)$ is small).

3 Experiments

3.1 Setup

Datasets. We evaluate our model on two public MSD datasets: (1) MMSD2.0 (Qin et al. 2023), a corrected version of the raw MMSD dataset (Cai, Cai, and Wan 2019) from Twitter (rebranded as X), which removes spurious cues and fixes unreasonable annotations, and (2) RedEval (Tang et al. 2024), which provides a test set from Reddit that we use as OOD data to assess the model generalizability. Detailed statistics of these two datasets are reported in Table 1.

[1]https://www.nltk.org/

[2]https://cloud.google.com/vision/docs/object-localizer

Dataset	Split	#Pos.	#Neg.	#Total	#Avg. Token
MMSD2.0	Train	9,576	10,240	19,816	13.42
	Validation	1,042	1,368	2,410	13.64
	Test	1,037	1,372	2,409	13.52
RedEval	Test	395	609	1,004	7.35

Table 1: Dataset statistics. "#Pos." and "#Neg." denote the number of positive and negative samples. "#Avg. Token" denotes the average number of words per sample.

Evaluation Metrics. Following previous works (Qin et al. 2023; Tang et al. 2024; Chen et al. 2024; Yuan et al. 2025), we adopt accuracy (Acc.), macro-average precision (P), recall (R) and F1 score to assess the model performance.

Comparison Models. We select a variety of MSD models, which are categorized as follows: (1) *Classification-based*: Multi-view CLIP (Qin et al. 2023), DGP (Ma et al. 2024b), CofiPara (Chen et al. 2024), and ESAM (Yuan et al. 2025); (2) *Generation-based*: LLaMA2-7B (Touvron et al. 2023), LLaVA-1.5-7B (Liu et al. 2024), Tang et al. (2024), LLaVA-1.5-7B + STaR (Zelikman et al. 2022), and LLaVA-1.5-7B + V-STaR (Hosseini et al. 2024) ; and (3) Reasoning-based: GPT-4o (Hurst et al. 2024), o3, Gemini 1.5 Pro (Team et al. 2024), and Gemini 2.5 Pro (Comanici et al. 2025).

Implementation Details. We employ LLaVA-1.5-7B (Liu et al. 2024) as primary LVLM. For further validation, we also conduct experiments with Qwen2-VL-7B (Wang et al. 2024) and Qwen2.5-VL-7B (Bai et al. 2025). Following previous works (Tang et al. 2024), we adopt LoRA (Hu et al. 2022) with a rank of 128 and alpha of 256 for LLaVA-1.5, and a rank of 64 with alpha 16 for Qwen2-VL. We employ DeepSpeed Zero2 (Rasley et al. 2020), maintaining a global batch size of 32 and a learning rate of $3e-5$. We run a total of 3 iterations for MMSD2.0, with 2 epochs of training per iteration. All experimental results are averaged over 5 runs.

3.2 Main Results

Table 2 summarizes the main results of the proposed S^3-MSD against baselines. For cost-efficiency, we further evaluate the reasoning-based models and cross-architecture generalizability of S^3-MSD on our custom MMSD2.0* subtest, which contains 500 instances (250 sarcastic, 250 non-sarcastic) randomly sampled from the test set. From Table 2 and Figure 4, we can obtain the following observations:

(1) S^3-MSD *outperforms all SOTA baselines across both in-distribution and OOD datasets, demonstrating its capability to effectively capture subtle inconsistencies in MSD*. Compared with LLaVA-1.5 which directly trains on golden labels, incorporating explanations in S^3-MSD achieves accuracy improvements of 4.12% and 3.88% on MMSD2.0 and RedEval, respectively. Furthermore, the enhanced generalizability from visual entity emphasis enables S^3-MSD to maintain robust performance under OOD scenarios.

(2) *The performance gain on RedEval is more notable, with* S^3-MSD *achieving a 2.98% accuracy improvement compared to 2.17% on MMSD2.0*. This can be attributed to the *self-focusing module*, which emphasizes sarcasm-related vi-

Model	w/ E?	MMSD2.0 (In-distribution)				RedEval (Out-of-distribution)			
		Acc. (%)	P (%)	R (%)	F1 (%)	Acc. (%)	P (%)	R (%)	F1 (%)
Multi-view CLIP (Qin et al. 2023)	✗	85.64	80.33	88.24	84.10	80.98	80.85	82.62	80.73
DGP (Ma et al. 2024b)	✗	87.21	87.10	86.48	86.75	-	-	-	-
CofiPara (Chen et al. 2024)	✗	85.70	85.96	85.55	85.89	-	-	-	-
ESAM (Yuan et al. 2025)	✗	85.87	83.12	86.05	84.56	-	-	-	-
LLaMA2-7B* (Touvron et al. 2023)	✗	84.68	84.40	84.94	84.53	81.38	80.47	80.60	80.53
LLaVA-1.5-7B (Liu et al. 2024)	✗	85.18	85.89	85.20	85.11	82.77	83.66	82.25	82.44
+ Tang et al. (2024)	✗	86.43	87.00	86.30	86.34	83.47	83.12	82.60	82.83
+ STaR (Zelikman et al. 2022)	✓	86.81	86.71	86.59	86.65	83.49	83.40	82.92	83.16
+ V-STaR (Hosseini et al. 2024) ($k = 3$)	✓	87.45	87.42	87.30	87.36	84.35	84.12	83.87	83.99
+ V-STaR (Hosseini et al. 2024) ($k = 5$)	✓	87.68	87.59	87.49	87.54	84.73	84.55	84.42	84.48
S^3-MSD (Ours)	✓	**88.69**	**88.74**	**88.43**	**88.57**	**85.98**	**85.62**	**85.06**	**85.34**

Table 2: Main results. "w/ E?" denotes whether the answer is provided with detailed explanations. "$*$" indicates that the model uses image captions as visual inputs. "k" refers to the number of sampled candidate explanations during inference. Best results are in bold. The improvements over all baselines are statistically significant with $p < 0.05$ under the t-test.

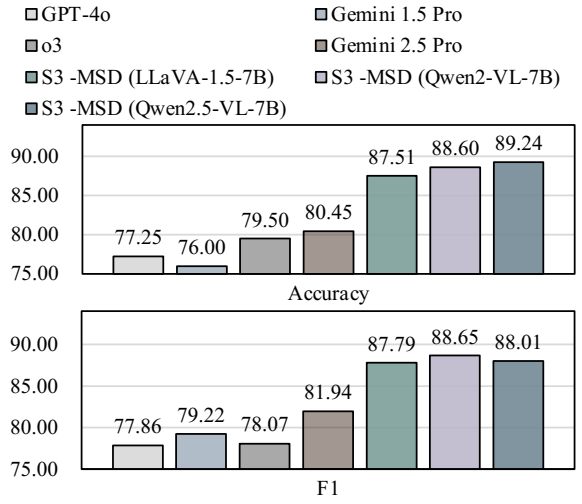


Figure 4: Comparison between S^3-MSD and other SOTA open-source and close-source LVLMs. Note that each model provides answers accompanied by detailed explanations.

sual entities and effectively mitigates over-reliance on textual information, thereby improving generalizability.

(3) S^3-MSD *is compatible with other LVLMs and outperforms untuned closed-source models.* The fine-tuned open-source model outperforms closed-source methods, particularly achieving an average accuracy improvement of round 10% on the MMSD2.0* test set. This demonstrates that task-specific, relatively smaller models can effectively outperform general-purpose, larger models on specialized tasks. Furthermore, the proposed model is compatible with different backbone architectures and significantly enhances the performance of its original counterpart by a large margin.

3.3 Explainability Analysis

Human Evaluation. To evaluate explainability, we conduct a human evaluation comparing our model's responses against the baseline V-STaR. We randomly sampled 100 instances, where three annotators judged whether the explanation from S^3-MSD was better (`Win`), worse (`Lose`), or of comparable quality (`Tie`). The final label for each instance was determined by majority vote. Cases of complete disagreement (*i.e.*, one vote for each category) were re-evaluated by the annotators to reach a consensus. As shown in Figure 5, the results indicate that: (1) Explanations from S^3-MSD were judged as superior or equal to V-STaR's in 81% of cases on MMSD2.0 and 85% on RedEval. (2) The higher win rate on the out-of-distribution RedEval dataset further highlights our model's strong generalization ability.

Evaluation on Sarcastic-only Benchmarks. We further benchmark S^3-MSD against SOTA methods on the MORE dataset (Desai, Chakraborty, and Akhtar 2022), which consists entirely of sarcastic samples with ground-truth explanations. To ensure a fair comparison with single-task models (*i.e.*, detection-only or explanation-only), we fine-tune our model on each dataset separately. Experiments on training with combined data are detailed in the Appendix.

As shown in Table 4, the proposed S^3-MSD fine-tuned on MORE achieves comparable results on METEOR (55.30) and BERT-Score (92.03), demonstrating superior fluency and semantic alignment. Furthermore, the proposed S^3-MSD unifies detection and explanation in a single framework, enhancing its applicability in real-world scenarios.

3.4 Ablation Study

As shown in Table 3, the ablation results validate the effectiveness of each component. Specifically, each component positively contributes to the performance of the proposed S^3-MSD. Additional ablation variants are presented below.

Necessity of Self-Calibrating. Since we utilize DPO in the visual modality to emphasize visual semantic entities, a natural question arises: why apply it to the textual modality as well, given that positive and negative samples are naturally generated through sampling? To explore this, we conducted controlled experiments by removing the self-calibrating module and replacing it with DPO. As observed in Figure 6, the reinforced version for the textual modality exhibits a significant performance drop across both datasets. We suspect this decline stems from two reasons: (1) MSD

Model	MMSD2.0 (In-distribution)				RedEval (Out-of-distribution)			
	Acc.(%)	P(%)	R(%)	F1(%)	Acc.(%)	P(%)	R(%)	F1(%)
S^3-MSD (Ours)	**88.69**	**88.74**	**88.43**	**88.57**	**85.98**	**85.62**	**85.06**	**85.34**
w/o *Self-Training*	86.65 (-2.04)	87.07 (-1.67)	86.60 (-1.83)	86.83 (-1.74)	84.12 (-1.86)	84.33 (-1.29)	83.44 (-1.62)	83.88 (-1.46)
w/o *Self-Calibrating*	87.64 (-1.05)	87.88 (-0.86)	87.65 (-0.78)	87.76 (-0.81)	85.14 (-0.84)	84.98 (-0.64)	84.49 (-0.57)	84.73 (-0.60)
w/o *Self-Focusing*	87.31 (-1.38)	87.55 (-1.19)	87.16 (-1.27)	87.35 (-1.22)	84.64 (-1.34)	84.47 (-1.15)	83.79 (-1.27)	84.13 (-1.21)

Table 3: Ablation study of key components.

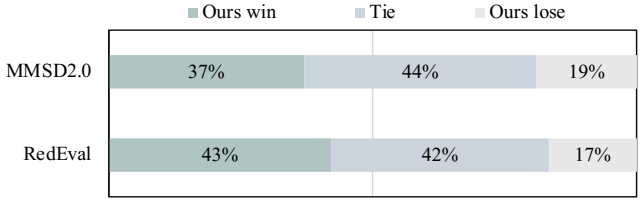


Figure 5: Human evaluation between the best baseline V-STaR and our proposed S^3-MSD model.

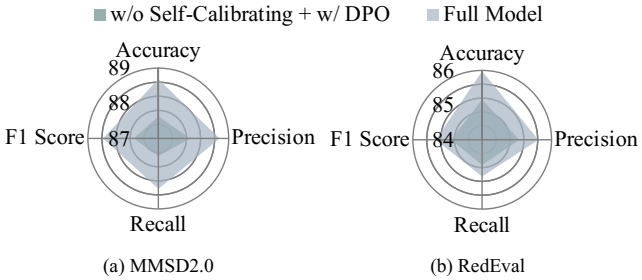
Model	B4	RL	MTR	BERT-Score
GPT-4o-mini (Hurst et al. 2024) + Zero-shot	2.04	20.68	26.45	86.42
GPT-4o-mini (Hurst et al. 2024) + One-shot	2.23	22.01	25.89	86.79
TEAM (Jing et al. 2023)	33.16	50.58	50.95	91.70
TURBO (Goel, Chauhan, and Akhtar 2025)	**35.26**	**53.12**	55.17	91.86
S^3-MSD (Ours)	34.25	52.37	**55.30**	**92.03**

Table 4: Explanation comparison on the MORE (Desai, Chakraborty, and Akhtar 2022) dataset. Results of GPT-4o-mini is from Goel, Chauhan, and Akhtar (2025). Best and second-best results are in bold and underlined, respectively. B4: BLEU-4; RL: ROUGE-L; MTR: METEOR.

involves complex semantic understanding, making conflicting reasoning particularly challenging, and (2) the diverse reasoning paths in textual modality make it difficult to determine a clear preferred response. Similar to the visual modality, exploring finer-grained textual preference optimization could be a promising direction for future research.

Generalizability on Metaphor Detection. To further evaluate the model generalizability, we conduct preliminary experiments on the multimodal metaphor detection (MMD) task, which requires determining the presence of metaphorical features and providing a classification result for a given set of cross-modal sample pairs. For a fair comparison with the baselines, we fine-tune BLIP2-2.7B (Li et al. 2023) using LoRA without careful hyperparameter selection.

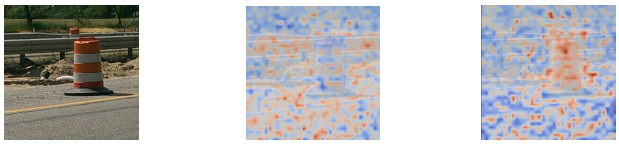
As shown in Table 5, S^3-MSD achieves superior results in terms of both accuracy and F1 score compared to baselines. Unlike the best baseline C4MMD (Xu et al. 2024) to employ CoT prompting in LVLMs to generate metaphorical cues, S^3-MSD refines metaphorical cues through iterative training. By progressively refining textual responses and emphasizing key visual features, S^3-MSD demonstrates promising results in deep semantic understanding.



(a) MMSD2.0 (b) RedEval

Figure 6: Comparison between our S^3-MSD and the reinforced version (DPO) for textual modality.

Text: *not even <num> miles from the house , ... and it 's so good to see michigan 's state flower blooming !*
Label: 1 (Sarcastic)
Image:

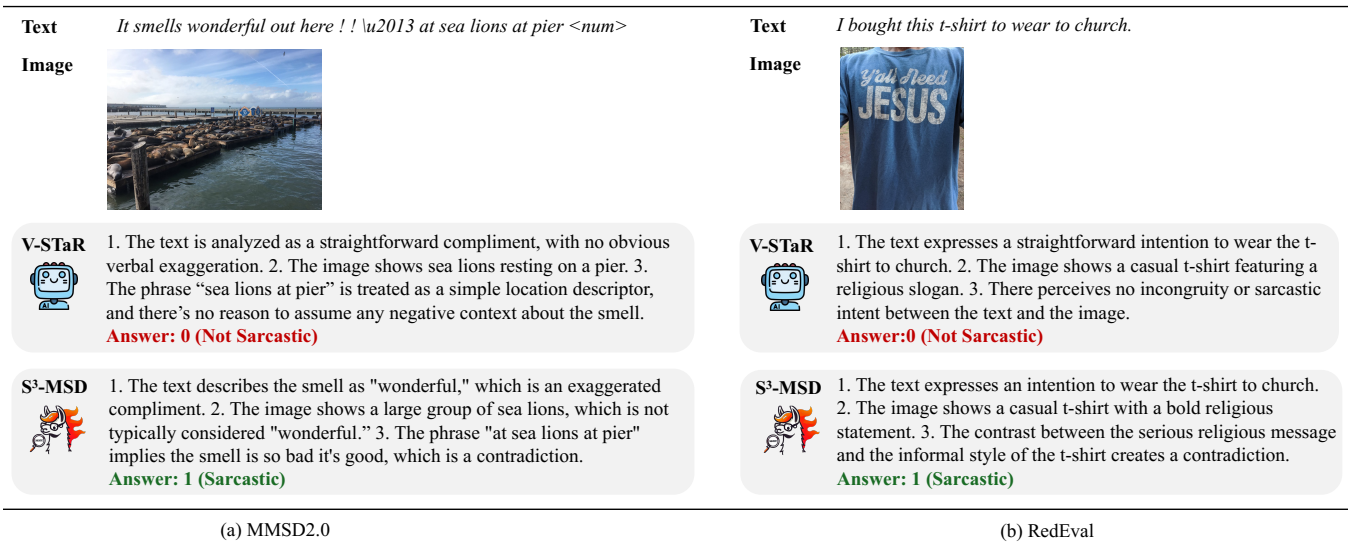


(a) Original (b) V-STaR (c) S^3-MSD

Figure 7: Visualization of the attention maps over the visual features between V-STaR and S^3-MSD. For each layer in the language model, we retrieve attention scores for all positions of the visual tokens, average them over all attention heads, and overlay corresponding heat maps with the image.

3.5 Qualitative Analysis

Visualization. Apart from theoretically analyzing how emphasizing visual entities can mitigate over-reliance on textual modality, we also visualize the attention map of visual features between the best baseline V-STaR and our S^3-MSD to intuitively demonstrate how our model operates on the visual modality. As shown in Figure 7, traffic barriers in the image are key factors in capturing sarcasm. S^3-MSD leverages entity-driven preference optimization to adjust its focus, ensuring that more attention is allocated to sarcasm-related visual entities. By prioritizing key entities, S^3-MSD effectively reduces reliance on textual information and enhances its ability to capture visual sarcasm cues.



(a) MMSD2.0 (b) RedEval

Figure 8: Case study between the proposed S^3-MSD and best baseline V-STaR on MMSD2.0 (a) and RedEval (b). Answers in red denote mispredictions, while the correct ones are in green. *(zoom-in for better view)*

Model	Acc.	P	R	F1
CLIP (Zhao et al. 2023)	75.05	60.83	**83.07**	70.23
BLIP2-2.7b (Li et al. 2023)	85.66	80.61	78.34	79.46
MultiCMET (Zhang et al. 2023)	85.66	82.69	75.25	78.79
C4MMD (Xu et al. 2024)	87.70	**83.33**	81.58	82.44
S^3-MSD (Ours)	**88.23**	83.26	82.10	**82.68**

Table 5: Performance comparison for multimodal metaphor detection on the MET-Meme dataset (Xu et al. 2022).

Case study. To intuitively verify how our model works, we present two cases comparing V-STaR and S^3-MSD on MMSD2.0 and RedEval, respectively. As depicted in Figure 8, V-STaR makes errors due to its reliance on text in case (a), where the word "wonderful" typically appears in non-sarcastic contexts. Similarly, in case (b), V-STaR incorrectly interprets the image when incorporating the text. In contrast, the proposed S^3-MSD leverages iterative explanation generation and calibration along with visual entity emphasis, to correctly capture cross-modal inconsistencies.

4 Related Work

Multimodal Sarcasm Detection. MSD has emerged as a critical research task for identifying cross-modal incongruity (Zhu et al. 2024a,b; Tian et al. 2023; Tang et al. 2024). Existing MSD methods can be broadly categorized as *classification-based* or *generation-based* (Zhu et al. 2025).

Classification-based methods design sophisticated fusion: early attention architectures (Pan et al. 2020); recent GNNs (Wu et al. 2020; Liang et al. 2022; Wei et al. 2024); knowledge enhancement (Liu, Wang, and Li 2022; Wei et al. 2025); and disentanglement of modalities (Qin et al. 2023). *Generation-based* approaches now leverage LVLMs: Tang et al. (2024) introduce CLIP-based (Radford et al. 2021) demonstration retrieval for in-context learning.

Despite promising results achieved, these works lack explainability and OOD robustness. Related sarcasm-explanation studies (Desai, Chakraborty, and Akhtar 2022; Jing et al. 2023) target sarcasm datasets but neglect detection, limiting real-world use. We first bridge detection and explanation for comprehensive sarcasm understanding.

Self-Training Frameworks. Self-training, which leverages a model's own outputs for iterative improvement, is a well-established paradigm (Deng et al. 2024; Wang, Li, and Lu 2024; Khan et al. 2024). A prominent example is STaR (Zelikman et al. 2022), where a model refines itself using its self-generated correct rationales. This concept has been extended by methods like rejection sampling for math reasoning (Yuan et al. 2023) and V-STaR (Xie et al. 2024), which employs a verifier trained via Direct Preference Optimization (DPO) (Rafailov et al. 2023) to rank and select the best of multiple generated explanations. Other alignment techniques like Reinforcement Learning from Human Feedback (RLHF) (Christiano et al. 2017) and various DPO applications (Ouali et al. 2024) are also widely used.

In this paper, we adopt a self-training framework to alleviate the scarcity of explanatory data in MSD. Furthermore, we introduce a self-calibrating module and a self-focusing module that iteratively refine imperfect explanations while enhancing the emphasis on salient visual entities.

5 Conclusion

This paper proposed S^3-MSD, a new LVLM for explainable and generalizable MSD. To develop S^3-MSD, we incorporate a *self-training paradigm* and a *self-calibrating mechanism* to enhance explainability, as well as a *self-focusing module* to improve generalizability in OOD scenarios. Experimental results demonstrate that S^3-MSD not only achieves SOTA performance in detection but also generates persuasive explanations. Method analysis including component ablation, cross-architecture comparisons, and cross-task evaluations reveals the potential of our model.

Acknowledgments

Bowen Xing was supported by the National Natural Science Foundation of China (Grant No. 62506033).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv:2502.13923*.

Cai, Y.; Cai, H.; and Wan, X. 2019. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *Proc. of ACL*, 2506–2515.

Chen, Z.; Lin, H.; Luo, Z.; Cheng, M.; Ma, J.; and Chen, G. 2024. CofiPara: A coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models. In *Proc. of ACL*, 9663–9687.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Proc. of NeurIPS*.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv:2507.06261*.

Deng, Y.; Lu, P.; Yin, F.; Hu, Z.; Shen, S.; Gu, Q.; Zou, J. Y.; Chang, K.-W.; and Wang, W. 2024. Enhancing large vision language models with self-training on image comprehension. *Proc. of NeurIPS*, 131369–131397.

Desai, P.; Chakraborty, T.; and Akhtar, M. S. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proc. of AAAI*, 10563–10571.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*, 4171–4186.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv:2010.11929*.

Farabi, S.; Ranasinghe, T.; Kanojia, D.; Kong, Y.; and Zampieri, M. 2024. A survey of multimodal sarcasm detection. In *Proc. of IJCAI*, 8020–8028.

Gibbs, R. W. 1986. On the psycholinguistics of sarcasm. *Journal of experimental psychology: general*, 3.

Goel, P.; Chauhan, D. S.; and Akhtar, M. S. 2025. Target-Augmented Shared Fusion-based Multimodal Sarcasm Explanation Generation. *arXiv:2502.07391*.

Hosseini, A.; Yuan, X.; Malkin, N.; Courville, A.; Sordoni, A.; and Agarwal, R. 2024. V-STaR: Training Verifiers for Self-Taught Reasoners. In *COLM*.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proc. of ICLR*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv:2410.21276*.

Jing, L.; Song, X.; Ouyang, K.; Jia, M.; and Nie, L. 2023. Multi-source Semantic Graph-based Multimodal Sarcasm Explanation Generation. In *Proc. of ACL*, 11349–11361.

Joshi, A.; Bhattacharyya, P.; and Carman, M. J. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 1–22.

Khan, Z.; BG, V. K.; Schulter, S.; Fu, Y.; and Chandraker, M. 2024. Self-training large language models for improved visual program synthesis with visual reinforcement. In *Proc. of CVPR*, 14344–14353.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. of ICML*, 19730–19742.

Li, J.; Pan, H.; Lin, Z.; Fu, P.; and Wang, W. 2021. Sarcasm detection with commonsense knowledge. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 3192–3201.

Liang, B.; Gui, L.; He, Y.; Cambria, E.; and Xu, R. 2024a. Fusion and discrimination: A multimodal graph contrastive learning framework for multimodal sarcasm detection. *IEEE Transactions on Affective Computing*.

Liang, B.; Lou, C.; Li, X.; Yang, M.; Gui, L.; He, Y.; Pei, W.; and Xu, R. 2022. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network. In *Proc. of ACL*, 1767–1777.

Liang, Z.; Xu, Y.; Hong, Y.; Shang, P.; Wang, Q.; Fu, Q.; and Liu, K. 2024b. A Survey of Multimodel Large Language Models. In *CAICE*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024. Improved baselines with visual instruction tuning. In *Proc. of CVPR*, 26296–26306.

Liu, H.; Wang, W.; and Li, H. 2022. Towards Multi-Modal Sarcasm Detection via Hierarchical Congruity Modeling with Knowledge Enhancement. In *Proc. of EMNLP*, 4995–5006.

Lou, C.; Liang, B.; Gui, L.; He, Y.; Dang, Y.; and Xu, R. 2021. Affective dependency graph for sarcasm detection. In *Proc. of SIGIR*, 1844–1849.

Ma, H.; He, D.; Wang, X.; Jin, D.; Ge, M.; and Wang, L. 2024a. Multi-Modal Sarcasm Detection Based on Dual Generative Processes. In *Proc. of IJCAI*, 2279–2287.

Ma, H.; He, D.; Wang, X.; Jin, D.; Ge, M.; and Wang, L. 2024b. Multi-modal sarcasm detection based on dual generative processes. In *Proc. of IJCAI*, 2279–2287.

Ouali, Y.; Bulat, A.; Martinez, B.; and Tzimiropoulos, G. 2024. Clip-dpo: Vision-language models as a source of preference for fixing hallucinations in lvlms. In *Proc. of ECCV*, 395–413.

Pan, H.; Lin, Z.; Fu, P.; Qi, Y.; and Wang, W. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Proc. of EMNLP Findings*, 1383–1392.

Qin, L.; Huang, S.; Chen, Q.; Cai, C.; Zhang, Y.; Liang, B.; Che, W.; and Xu, R. 2023. MMSD2.0: Towards a Reliable Multi-modal Sarcasm Detection System. In *Proc. of ACL Findings*, 10834–10845.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *Proc. of ICML*, 8748–8763.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. *Proc. of NeurIPS*, 53728–53741.

Rasley, J.; Rajbhandari, S.; Ruwase, O.; and He, Y. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proc. of KDD*, 3505–3506.

Tang, B.; Lin, B.; Yan, H.; and Li, S. 2024. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In *Proc. of NAACL*, 1732–1742.

Team, G.; Georgiev, P.; Lei, V. I.; Burnell, R.; Bai, L.; Gulati, A.; Tanzer, G.; Vincent, D.; Pan, Z.; Wang, S.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv:2403.05530*.

Tian, Y.; Xu, N.; Zhang, R.; and Mao, W. 2023. Dynamic Routing Transformer Network for Multimodal Sarcasm Detection. In *Proc. of ACL*, 2468–2480.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv:2409.12191*.

Wang, T.; Li, S.; and Lu, W. 2024. Self-Training with Direct Preference Optimization Improves Chain-of-Thought Reasoning. In *Proc. of ACL*, 11917–11928.

Wei, Y.; Yuan, S.; Zhou, H.; Wang, L.; Yan, Z.; Yang, R.; and Chen, M. 2024. Gˆ 2SAM: Graph-Based Global Semantic Awareness Method for Multimodal Sarcasm Detection. In *Proc. of AAAI*, 9151–9159.

Wei, Y.; Zhou, H.; Yuan, S.; Chen, M.; Shi, H.; Jia, Z.; Wang, L.; and He, X. 2025. DeepMSD: Advancing Multimodal Sarcasm Detection through Knowledge-augmented Graph Reasoning. *TCSVT*.

Wen, C.; Jia, G.; and Yang, J. 2023. Dip: Dual incongruity perceiving network for sarcasm detection. In *Proc. of CVPR*, 2540–2550.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 4–24.

Xie, Y.; Li, G.; Xu, X.; and Kan, M.-Y. 2024. V-DPO: Mitigating Hallucination in Large Vision Language Models via Vision-Guided Direct Preference Optimization. In *Proc. of EMNLP Findings*, 13258–13273.

Xu, B.; Li, T.; Zheng, J.; Naseriparsa, M.; Zhao, Z.; Lin, H.; and Xia, F. 2022. Met-meme: A multimodal meme dataset rich in metaphors. In *Proc. of SIGIR*, 2887–2899.

Xu, N.; Zeng, Z.; and Mao, W. 2020. Reasoning with Multimodal Sarcastic Tweets via Modeling Cross-Modality Contrast and Semantic Association. In *Proc. of ACL*, 3777–3786.

Xu, Y.; Hua, Y.; Li, S.; and Wang, Z. 2024. Exploring Chain-of-Thought for Multi-modal Metaphor Detection. In *Proc. of ACL*, 91–101.

Yuan, S.; Wei, Y.; Zhou, H.; Xu, Q.; Chen, M.; and He, X. 2025. Enhancing Semantic Awareness by Sentimental Constraint with Automatic Outlier Masking for Multimodal Sarcasm Detection. *IEEE Transactions on Multimedia*.

Yuan, Z.; Yuan, H.; Li, C.; Dong, G.; Lu, K.; Tan, C.; Zhou, C.; and Zhou, J. 2023. Scaling relationship on learning mathematical reasoning with large language models. *arXiv:2308.01825*.

Zelikman, E.; Wu, Y.; Mu, J.; and Goodman, N. 2022. Star: Bootstrapping reasoning with reasoning. *Proc. of NeurIPS*, 15476–15488.

Zhang, D.; Yu, J.; Jin, S.; Yang, L.; and Lin, H. 2023. Multicmet: A novel chinese benchmark for understanding multimodal metaphor. In *Proc. of EMNLP Findings*, 6141–6154.

Zhang, F.; Cheng, Z.; Deng, C.; Li, H.; Lian, Z.; Chen, Q.; Liu, H.; Wang, W.; Zhang, Y.-F.; Zhang, R.; et al. 2025a. Mme-emotion: A holistic evaluation benchmark for emotional intelligence in multimodal large language models. *arXiv*.

Zhang, F.; Li, H.; Qian, S.; Wang, X.; Lian, Z.; Wu, H.; Zhu, Z.; Gao, Y.; Li, Q.; Zheng, Y.; et al. 2025b. Rethinking Facial Expression Recognition in the Era of Multimodal Large Language Models: Benchmark, Datasets, and Beyond. *arXiv*.

Zhang, Y.; Chen, Q.; Zhou, J.; Wang, P.; Si, J.; Wang, J.; Lu, W.; and Qin, L. 2024. Wrong-of-Thought: An Integrated Reasoning Framework with Multi-Perspective Verification and Wrong Information. In *Proc. of EMNLP Findings*, 6644–6653.

Zhao, B.; Zhang, A.; Watson, B.; Kearney, G.; and Dale, I. 2023. A review of vision-language models and their performance on the hateful memes challenge. *arXiv:2305.06159*.

Zhou, Y.; Fan, Z.; Cheng, D.; Yang, S.; Chen, Z.; Cui, C.; Wang, X.; Li, Y.; Zhang, L.; and Yao, H. 2024. Calibrated Self-Rewarding Vision Language Models. In *Proc. of NeurIPS*.

Zhu, Z.; Cheng, X.; Chen, Z.; Chen, Y.; Zhang, Y.; Wu, X.; Zheng, Y.; and Xing, B. 2024a. InMu-Net: advancing multimodal intent detection via information bottleneck and multisensory processing. In *ACM MM*.

Zhu, Z.; Zhang, F.; Zhang, Y.; Sun, J.; Huang, Z.; Long, Q.; Xing, B.; and Wu, X. 2025. A Survey on Multi-modal Intent Recognition: Recent Advances and New Frontiers. In *EMNLP*.

Zhu, Z.; Zhuang, X.; Zhang, Y.; Xu, D.; Hu, G.; Wu, X.; and Zheng, Y. 2024b. Tfcd: Towards multi-modal sarcasm detection via training-free counterfactual debiasing. In *IJCAI*.