CFMS: Towards Explainable and Fine-Grained Chinese Multimodal Sarcasm Detection Benchmark

Anonymous ACL submission

Abstract

Multimodal sarcasm detection has recently garnered significant attention. However, existing benchmarks suffer from coarse-grained annotations and limited cultural coverage, which hinder research into fine-grained semantic understanding. To address this, we construct **CFMS**, the first fine-grained multimodal sarcasm dataset tailored for Chinese social media. It comprises 2,796 high-quality image-text pairs and provides a triple-level annotation framework: sarcasm identification, target recognition, and explanation generation. We find that the fine-grained explanation annotations effectively guide AI in generating images with explicit sarcastic intent. Furthermore, we curate a high-consistency parallel Chinese-English metaphor subset (200 entries each), revealing significant limitations of current models in metaphoric reasoning. To overcome the constraints of traditional retrieval methods, we propose a Reinforcement Learning-augmented In-Context Learning strategy (**PGDS**) to dynamically optimize exemplar selection. Extensive experiments demonstrate that CFMS provides a solid foundation for building reliable multimodal sarcasm understanding systems, and the PGDS method significantly outperforms existing baselines on key tasks. Our data and code are available at https://anonymous.4open.science/r/CFMS-E8F9.

1 Introduction

Sarcasm is a sophisticated linguistic phenomenon where literal meanings deviate from true intents to convey mockery or criticism (Camp, 2012). In the multimodal era, sarcasm frequently manifests through the conflict between text and imagery (Schifanella et al., 2016). As illustrated in Figure 1, the phrase "Install fully consistent heads" appears superficially neutral, yet its pairing with a mechanized conveyor belt reveals a deep-seated social critique. Resolving such **culturally-grounded semantic inversion** constitutes a primary challenge for automated systems.

Figure 1: A representative instance of Chinese multimodal sarcasm from the CFMS dataset. The neutral text "装上全部一致的脑袋" (lit. "Install fully consistent heads") contrasts with the visual of a conveyor belt, satirizing ideological conformity via culturally-grounded semantic inversion.

While multimodal sarcasm detection (MSD) has progressed with benchmarks like MMSD2.0 (Qin et al., 2023), existing research still faces three primary limitations: (1) **Cultural Bias:** Most datasets are English-centric, failing to capture unique Chinese sarcastic forms like "阴阳怪气" (passive-aggression). (2) **Coarse-grained Annotation:** Binary classification tasks often ignore the *target* and *construction mechanism*, limiting analytical depth. (3) **Shallow Heuristics:** Models frequently rely on surface cues (e.g., hashtags) rather than integrated multimodal reasoning (Cai et al., 2019).

To address these gaps, we propose **CFMS** (Chinese Fine-grained Multimodal Sarcasm), a benchmark oriented toward real-world Chinese social media. CFMS shifts the focus from classification to interpretation via a **triple-level annotation** framework: *sarcasm identification*, *target recognition*, and *explanation generation*.

Furthermore, we explore the **Transferability** of CFMS beyond detection. Inspired by the synergy

between sarcasm and metaphor in persuasive communication (Van Mulken et al., 2005), we demonstrate that our reason-level explanations can serve as *structured instruction signals* for AI-Generated Content (AIGC). In domains like creative advertising, CFMS enables Text-to-Image models to manifest implicit rhetorical intents that surface-level prompts fail to evoke (Oppenlaender, 2024), bridging the gap between understanding irony and generating it.

Regarding methodology, we evaluate various MLLMs and propose **Policy-Guided Demonstration Selection (PGDS)**. PGDS is a reinforcement-learned strategy that dynamically optimizes the selection of in-context exemplars based on task-level feedback, effectively improving fine-grained sarcasm understanding without modifying model parameters.

Our contributions are as follows:

- We provide an in-depth analysis of the limitations in current MSD benchmarks regarding annotation granularity and cultural coverage.
- We construct **CFMS**, the first fine-grained Chinese multimodal sarcasm benchmark with a triple-annotation system to facilitate interpretable sarcasm understanding.
- We propose **PGDS**, a reinforcement-learned strategy for dynamic exemplar selection that significantly enhances model performance in complex scenarios.
- We validate the **transferability** of fine-grained sarcasm explanations, demonstrating their value in guiding creative AI content generation.

2 Related Work

Multimodal sarcasm detection originated with the work of Schifanella et al. (2016), who first attempted to identify sarcasm by concatenating visual and textual features. Cai et al. (2019) released the MMSD dataset, which became one of the most widely used benchmarks in the field. However, this dataset has been noted for containing significant spurious cues and annotation inconsistencies.

To address these reliability issues, Qin et al. (2023) introduced MMSD2.0, which improved data quality by removing noisy samples and re-annotating unreasonable instances. From a different perspective, Castro et al. (2019) constructed the MUStARD dataset using video clips from popular TV shows such as *Friends* and *The Big Bang Theory*, providing richer contextual information. Nevertheless, MUStARD is limited by its small scale and single cultural background.

In terms of methodology, early approaches evolved from simple feature concatenation to hierarchical fusion models based on attention mechanisms. Liang et al. (2022) proposed a cross-modal graph convolutional network to model atomic-level and composition-level consistencies across modalities. Recently, Yue et al. (2024) introduced SarcNet, the first bilingual multimodal sarcasm dataset; however, it does not yet address fine-grained tasks such as target recognition or explanation generation. In contrast, CFMS focuses on the Chinese social media context and is the first to incorporate target recognition and explanation generation, pushing the field towards interpretability. Notably, a recent survey by Farabi et al. (2024) emphasizes that high-quality dataset construction and fine-grained task design are paramount for future research—a principle that our work strictly adheres to.

3 Dataset Construction

3.1 Preliminaries: Cognitive Framework for Sarcasm

The design of the CFMS dataset is grounded in the cognitive mechanisms of sarcasm. Human comprehension of sarcasm stems from complex processes involving semantic conflict, pragmatic reasoning, and cultural background knowledge (Gibbs, 2017). While MLLMs have shown promise in general sentiment analysis, they often struggle with scenarios relying on deep semantic conflict reasoning—areas where humans maintain a significant advantage by evaluating the "sarcastic effect" (Lu et al., 2023).

To bridge this gap, we adopt a structured expression of sarcasm explanation. Unlike previous benchmarks that focus solely on identification, we follow the insights of Konstantinidou et al. (2025) to capture semantic cues such as visual anomalies and emotional misalignment. In CFMS, we translate these human cognitive processes into computable, fine-grained annotations. This framework not only provides an interpretable path for models but also acts as structured guidance for generative tasks, moving from "understanding" to "generating" sarcastic intent.

3.2 Sarcasm Data Collection

We select a major Chinese social media platform as our primary data source, whose young user base frequently employs domain-specific sarcastic expres-

sions like “阴阳怪气” (passive-aggressive speech) and “内涵” (veiled allusions). This provides a sharp contrast to existing English-centric datasets like MMSD (Cai et al., 2019) and MUStARD (Castro et al., 2019), allowing us to capture the unique sarcastic culture of Chinese social media.

Regarding data acquisition, we adopt a directed crawling strategy. Unlike the Twitter API method used in MMSD, we utilize a distributed crawler framework to perform keyword-based searches using terms such as “sarcasm,” “irony,” “connotation images,” and “passive-aggression,” obtaining approximately 10,000 raw entries. To resolve the modality misalignment issue (where one text is paired with multiple images), we employ PaddleOCR to extract embedded text within images and perform semantic fusion with external captions, constructing strongly correlated “image-text” pairs.

Then, A triple-stage purification pipeline was implemented: (1) **pHash Deduplication:** Removing duplicate samples with a similarity score $> 90\%$. (2) **Commercial Filter:** Filtering advertisements by detecting watermark area ($> 15\%$) and commercial logos. (3) **Low-Resolution Filter:** Discarding images with pixels $< 512 \times 512$. After purification, approximately 3,500 candidate samples remained.

For the annotation process, we designed an “LLM Pre-annotation + Human Verification” pipeline inspired by the **human-in-the-loop (HITL)** paradigm (Munro, 2021), which leverages the synergy between automated efficiency and human expertise (Wang et al., 2021; Gilardi et al., 2023). The workflow (illustrated in Figure 2) follows three steps: (1) **GPT-4o Pre-annotation** using a structured prompt (logic: “Explicit Element Recording $\rightarrow$ Conflict Detection $\rightarrow$ Counter-Hypothesis Verification”); (2) **Human Verification** by three groups of graduate students in computational linguistics; (3) **Interactive Error Correction** via an integrated GPT-4o API for real-time feedback. The custom-developed annotation interface is illustrated in Appendix A.3.

The annotation achieved a Kappa coefficient of 0.69 (“substantial agreement”), comparable to MUStARD (Castro et al., 2019). The BLEU-4 consistency for the explanation annotation reached 0.68, while the Kappa was 0.63, indicating high annotation reliability. The final CFMS dataset contains 2,796 samples. The text length statistics are summarized in Table 1, and the dataset split is detailed in Table 2.

As shown in the sarcasm target word cloud (Figure 3), targets are categorized into five types: *Social Phenomena* (41%), *Individual Behavior* (23%), *Interpersonal Relations* (17%), *Institutional Rules* (12%), and *Others* (7%). Table 3 presents three representative samples with their corresponding fine-grained annotations.

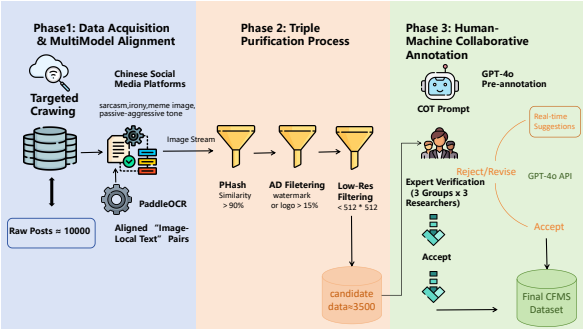


Figure 2: The human-in-the-loop annotation pipeline for the CFMS dataset, integrating GPT-4o Pre-annotation, Human Verification, and Interactive Error Correction.

Category	Avg. Length (Chars)
Text Context	18.17
Sarcasm Target	5.44
Sarcasm Explanation	61.69

Table 1: Text length distribution in CFMS (N=2,796).

3.3 Sarcastic Metaphor Data Construction

The relationship between metaphor and sarcasm is a cornerstone of linguistic research. Conceptual Metaphor Theory (Gibbs, 2017) posits that metaphors are fundamental to human cognition, and sarcasm is often built upon metaphoric foundations. Given the high subjectivity of metaphors, the initial Fleiss’ Kappa of our raw annotation was only around 0.15.

Furthermore, to facilitate cross-linguistic studies, we leveraged a previously collected English metaphor dataset sourced from Twitter archives. This subset, which originally only contained binary metaphor labels, was re-annotated using our triple-level framework to provide a high-quality parallel English reference (Table 5).

Comparing the English target word cloud (Figure 4) with the Chinese one, we observe that English sarcasm targets are more concentrated on **public policy** and **institutional efficiency**, whereas Chinese targets frequently focus on **social etiquette** and **workplace culture**, reflecting distinct cultural nuances in rhetorical expression.

Split	Positive	Negative	Total
Training	872	1,084	1,956
Validation	187	233	420
Test	187	233	420
Total	**1,246**	**1,550**	**2,796**

Table 2: Statistics and distribution of the CFMS dataset.

Figure 3: Word cloud of sarcasm targets in the Chinese CFMS dataset. Larger fonts indicate more frequently identified targets, such as "social phenomena" and "individual behavior."

3.4 Explanation-Driven AI Generation Test Set

We utilized CFMS annotations to guide `gemini-3-pro-image-preview` in generating 100 images. Two annotators evaluated whether the generated content achieved the intended sarcastic effect. As shown in Table 6, an average of 76% of the images successfully conveyed the intended sarcasm, confirming the value of CFMS labels as structured instructional signals for controllable generation. Failure cases (e.g., "Workplace PUA") reveal current limitations in reifying abstract cultural metaphors. Qualitative examples of AI-generated sarcastic images can be found in Appendix A.1.

4 Approach

4.1 Task Definition

CFMS comprises three progressively hierarchical subtasks designed to evaluate model capabilities across identification, localization, and explanation. We define the input as a multimodal pair $x = (I, T)$, where I represents the image and T represents the text.

Task 1: Sarcasm Identification. This task aims to determine whether the input contains sarcastic intent. We define it as a mapping function f_{det} that

Figure 4: Word cloud of sarcasm targets in the English parallel subset. It reveals a higher concentration of targets related to public policy and institutional efficiency compared to the Chinese subset.

maps the multimodal input to a binary label space:

$$f_{\text{det}} : I \times T \rightarrow \{0, 1\} \tag{1}$$

where 0 denotes non-sarcasm and 1 denotes sarcasm.

Evaluation Metrics: We employ a standard set of classification metrics $\mathcal{M}_{\text{cls}} = \{\text{Accuracy, Precision, Recall, F1}\}$. For the test set D_{test}, the discrepancy between the predicted value $\hat{y} = f_{\text{det}}(x)$ and the ground truth y is calculated.

Task 2: Sarcasm Target Recognition. Conditioned on identifying a sarcastic sample ($y = 1$), this task aims to localize the target of the sarcasm. It is defined as a mapping from the input to a target set V:

$$f_{\text{target}} : (I \times T \mid y = 1) \rightarrow t \in V \tag{2}$$

where t is the identified specific sarcastic target.

Evaluation Metrics: Given the open-ended nature of generation, we define a semantic consistency function $S(t_{\text{pred}}, t_{\text{gold}}) \in \{0, 1\}$. The accuracy is defined as:

$$\text{Acc}_{\text{target}} = \frac{1}{|D_{\text{pos}}|} \sum_{x \in D_{\text{pos}}} S(f_{\text{target}}(x), t_{\text{gold}}) \tag{3}$$

where D_{pos} denotes the subset of sarcastic (positive) instances in the test set, as target recognition is only evaluated on samples with valid sarcastic intent. $|D_{\text{pos}}|$ represents the cardinality of this subset. The consistency function S is implemented via GPT-4o-based semantic matching to account for linguistic variations in generated outputs.

Task 3: Sarcasm Explanation Generation. This task requires the model to generate a natural

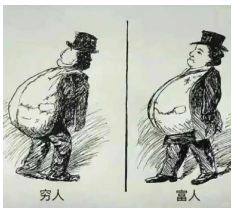
Case 1: Modern Human Activity	
	Text: 现代人类活动(Modern Human Activity) **Target**: 手机依赖(Mobile device dependency) **Explanation**: 通过人与狗在相似场景中的对比，讽刺人类对手机及充电设备的过度依赖。反映了科技在提供便利的同时，也带来了生理与精神上的束缚。 *(Translation: By contrasting humans and dogs in similar scenarios, this sample satirizes excessive reliance on mobile devices, reflecting how technology limits freedom while providing convenience.)*
Case 2: The Role Model	
	Text: 《榜样》(《The Role Model》) **Target**: 家长榜样作用缺失(Lack of parental role modeling) **Explanation**: 左边的家长使用手机，孩子随之模仿，右边的家长在读书，孩子也读书。通过对比揭示了一些家长以身作则不足的问题。 *(Translation: This case contrasts two parental behaviors—using phones vs. reading—and the child's subsequent imitation, revealing the failure of parental role modeling in fostering habits.)*
Case 3: Rich vs. Poor	
	**Text**: 穷人富人(Rich Poor) **Target**: 贫富差距(Wealth disparity) **Explanation**: 通过对比穷人负重弯曲的脊背与富人鼓胀的肚子，反映富人通过压榨穷人积累财富的现象，批判社会的贫富不均。 *(Translation: By comparing the hunched back of the poor carrying heavy loads with the bloated stomach of the rich, it reflects the accumulation of wealth through exploitation and criticizes social inequality.)*

Table 3: Fine-grained annotations of samples from the CFMS dataset.

Type (Chinese)	Positive	Negative
Metaphor	73	127
Sarcasm	71	129

Table 4: Distribution of the Chinese Sarcasm-Metaphor subset (N=200).

Type (English)	Positive	Negative
Metaphor	100	100
Sarcasm	100	100

Table 5: Distribution of the English parallel subset (N=200).

language text elucidating the construction mechanism of the sarcasm. The generation function is defined as follows:

$$f_{\text{exp}} : (I \times T \mid y = 1) \rightarrow E = (w_1, w_2, \ldots, w_n) \tag{4}$$

where E is the sequence of the explanation text.

Evaluation Metrics: We utilize a set of text generation quality metrics $\mathcal{M}_{\text{gen}} = \{\text{BLEU-4, BERTScore}\}$ to measure the semantic similarity between the generated explanation $\hat{E}$ and the human-annotated reference E_{ref}.

Annotator	Sarcastic Effect Met	Ratio (%)
Annotator A	70	70.0
Annotator B	82	82.0
Average	**76**	**76.0**

Table 6: Evaluation of AI-generated images based on CFMS annotations (N=100).

4.2 Policy-Guided Demonstration Selection

Traditional similarity-based retrieval methods often fall short when handling tasks like sarcasm detection that require deep semantic reasoning. To address this, we propose **Policy-Guided Demonstration Selection (PGDS)**, a reinforcement-learned strategy that dynamically optimizes exemplar selection to improve fine-grained sarcasm understanding. The comprehensive workflow is outlined in Algorithm 1.

Framework Design: Given a query sample $x_q = (I_q, T_q)$, where I_q is the image and T_q is the text, we first obtain a joint representation using a BGE encoder f_{BGE} (Xiao et al., 2024) and a CLIP encoder f_{CLIP} (Radford et al., 2021):

$$h_q = \text{Concat}(f_{\text{BGE}}(T_q), f_{\text{CLIP}}(I_q)) \tag{5}$$

Based on this representation, we retrieve the top-50 candidate samples from the training set $\mathcal{D}_{\text{train}}$:

$$\mathcal{C}_q = \{x_i \mid x_i \in \mathcal{D}_{\text{train}}, \text{sim}(h_q, h_i) \in \text{Top-50}\} \tag{6}$$

where $\text{sim}(h_q, h_i)$ denotes the cosine similarity.

The core innovation lies in the introduction of a lightweight policy network π_θ, parameterized as a two-layer MLP. Taking the joint embedding of the query and candidates as input, it outputs the selection probabilities:

$$\begin{aligned} p_i &= \pi_\theta(h_q, h_i) \\ &= \text{Softmax}(W_2\sigma(W_1[h_q; h_i] + b_1) + b_2) \end{aligned} \tag{7}$$

where σ is the ReLU activation function and $[;]$ denotes vector concatenation. After calculating the weights for each candidate, we employ a probability-based sampling mechanism to select k representative examples: $\mathcal{S}_q = \text{SampleTopK}(\{p_i\}_{i=1}^{50}, k)$. These examples are combined with the query into a structured prompt: $P = \text{Template}(\mathcal{S}_q, x_q)$. The multimodal model f_{MLLM} then generates a prediction: $\hat{y} = f_{\text{MLLM}}(P)$.

Reward Design and Optimization: Predictions are evaluated through a multi-dimensional reward function:

$$R = w_1 \cdot R_{\text{format}} + w_2 \cdot R_{\text{cls}} + w_3 \cdot R_{\text{target}} + w_4 \cdot R_{\text{exp}} \tag{8}$$

where $R_{\text{format}} \in \{0, 1\}$ assesses the output format compliance, $R_{\text{cls}} \in \{0, 1\}$ evaluates classification accuracy, and R_{target} and R_{exp} utilize BERTScore to measure the semantic quality of the identified target and generated explanation, respectively. The weights w_i are tuned via the validation set such that $\sum w_i = 1$.

The policy network is optimized using the REINFORCE algorithm (Williams, 1992), with the objective function: $J(\theta) = \mathbb{E}_{\mathcal{S}_q \sim \pi_\theta}[R(\mathcal{S}_q)]$. To reduce variance, gradient estimation incorporates a moving average baseline b_t:

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} (R_i - b_t) \nabla_\theta \log \pi_\theta(\mathcal{S}_q^{(i)} \mid h_q^{(i)}) \tag{9}$$

where $b_t = \gamma b_{t-1} + (1-\gamma) R_{t-1}$ is the exponential moving average with $\gamma = 0.9$.

Compared to static retrieval, our dynamic strategy demonstrates advantages in complex sarcastic scenarios. Particularly for samples involving metaphors, the policy network identifies examples that may have low surface similarity but are highly relevant in deep semantic structure. This adaptability is crucial for handling the diverse sarcastic expressions prevalent in Chinese social media.

Algorithm 1 PGDS: Policy-Guided Demonstration Selection

Require: Training set $\mathcal{D}_{\text{train}}$, Query x_q, MLLM f_{MLLM}, Policy π_θ, Exemplar size k
Ensure: Selected demonstrations $\mathcal{S}_q$ for inference

```
h_q ← Encode(x_q) {BGE for text, CLIP for image}
C_q ← TopK({sim(h_q, h_i) | x_i ∈ D_train}, 50)
for each candidate x_i ∈ C_q do
    p_i ← π_θ(h_q, h_i) {Compute selection probability}
end for
S_q ← Sample(C_q, {p_i}, k) {Probabilistic sampling}
P ← Template(S_q, x_q) {Construct structured prompt}
ŷ ← f_MLLM(P) {Generate interpretation}
R ← CalculateReward(ŷ, y_gold) {Format, Cls, BERTScore}
if in training phase then
    b_t ← γb_{t-1} + (1 − γ)R {Update baseline}
    θ ← θ + α(R − b_t)∇_θ log π_θ(S_q | h_q) {Policy gradient update}
end if

return S_q
```

5 Experiments

5.1 Experimental Setup

Baselines: We evaluate a diverse set of state-of-the-art MLLMs, including: (1) **Closed-source models:** GPT-4o (Hurst et al., 2024) and Gemini-2.5-Flash (Comanici et al., 2025). (2) **Open-source models:** Qwen2.5-VL-7B-Instruct (Bai et al., 2025), InternVL2.5-8B (Chen et al., 2025), and LLaVA-1.5-7B (Liu et al., 2023). Note that LLaVA-1.5-7B is excluded from 1-shot experiments as it does not natively support interleaved multi-image inputs.

Learning Paradigms: To explore different adaptation strategies, we implement two paradigms:

- **Parameter-Efficient Fine-Tuning (PEFT):** We employ LoRA (Hu et al., 2022) to adapt models by introducing low-rank matrices A and B to approximate the weight update:

$$h = Wx + BAx \tag{10}$$

  This allows the model to learn sarcastic patterns with minimal computational cost.

- **Retrieval-Based In-Context Learning (ICL):** We adopt an ICL framework that

Model	Method	Acc ↑	P	R	F1 ↑	Target Acc ↑	BLEU-4 (%)	BERTScore (%)
				CLOSED-SOURCE				
GPT-4o	Zero-shot	70.24	60.67	96.30	74.44	71.03	8.07	72.81
	Random 1-shot	76.36	70.08	86.41	77.39	61.17	8.39	72.59
	RAG 1-shot	77.62	68.20	94.18	79.10	73.02	**10.60**	**73.84**
Gemini-2.5-Flash	Zero-shot	71.19	60.31	95.24	74.84	**82.54**	6.30	72.41
	Random 1-shot	75.71	67.61	88.36	76.61	76.19	6.88	72.55
	RAG 1-shot	75.00	65.79	92.59	76.92	80.42	7.34	73.03
				OPEN-SOURCE				
Qwen2.5-VL-7B-Instruct	Zero-shot	70.71	61.64	73.54	69.33	32.54	7.29	72.04
	Random 1-shot	71.25	64.54	80.16	71.50	38.62	8.24	71.83
	RAG 1-shot	76.88	71.08	81.94	76.13	45.67	8.18	71.92
	PGDS (Ours)	78.01	76.92	75.76	76.34	48.68	8.10	71.52
InternVL2.5-8B	Zero-shot	68.48	59.67	93.72	72.91	40.47	6.77	70.72
	Random 1-shot	69.29	60.10	94.44	73.46	48.15	8.33	72.31
	RAG 1-shot	71.05	62.11	96.15	75.47	46.15	8.58	71.69
	PGDS (Ours)	74.76	66.27	89.42	76.13	50.89	8.76	72.30

Table 7: Main results on the CFMS dataset (excluding LoRA fine-tuning). **Zero-shot**, **Random 1-shot**, **RAG 1-shot** are baselines; **PGDS (Ours)** is our proposed dynamic demonstration selection strategy. Best results per column (across all models/methods) are in **bold**. "-" indicates no corresponding experimental results.

Model	Method	Acc ↑	P	R	F1 ↑	Target Acc ↑	BLEU-4 (%)	BERTScore (%)
Qwen2.5-VL-7B-Instruct	LoRA FT	82.30	81.22	78.61	79.89	49.73	7.51	71.24
InternVL2.5-8B	LoRA FT	**83.25**	79.70	83.96	**81.77**	53.48	9.35	72.36
LLaVA-1.5-7B	LoRA FT	80.95	**84.71**	70.37	76.88	36.51	5.71	69.32

Table 8: Experimental results of LoRA fine-tuning (LoRA FT) on the CFMS dataset. Best results per column are in **bold**.

retrieves the most similar samples s_i from D_{train} to construct an augmented prompt:

$$P = [s_1 \oplus \cdots \oplus s_k \oplus x_q] \tag{11}$$

The similarity is determined by the cosine distance of textual features:

$$\text{sim}(x_q, x_i) = \frac{\text{Enc}(T_q) \cdot \text{Enc}(T_i)}{\|\text{Enc}(T_q)\| \cdot \|\text{Enc}(T_i)\|} \tag{12}$$

Implementation Details: For LoRA, we set $r = 16, \alpha = 32$, targeting all linear modules with a 5×10^{-5} learning rate over 10 epochs. For ICL, we employ a multimodal retrieval strategy where `bge-large-zh-v1.5` is used for textual semantic encoding and `clip-vit-base-patch32` is used for image feature extraction. These multimodal embeddings are indexed using FAISS to facilitate efficient and accurate demonstration retrieval.

Evaluation Metrics: We report Accuracy (Acc), Precision (P), Recall (R), and F1-score for identification. For target recognition, GPT-4o serves as a semantic judge. Explanation generation is evaluated via BLEU-4 and BERTScore-F1.

5.2 Main Results and Analysis

Table 7 and Table 8 summarize model performance across different paradigms.

Performance of ICL Strategies: As shown in Table 7, 1-shot learning generally improves performance over zero-shot. While Random 1-shot provides a baseline, RAG 1-shot further leverages semantic similarity to enhance reasoning. However, closed-source models like Gemini-2.5-Flash show high sensitivity to demonstration quality, where Random 1-shot leads to a drop in Target Acc (from 82.54% to 76.19%).

The Superiority and Efficiency of PGDS: Our proposed **PGDS** consistently outperforms Random and RAG-based 1-shot baselines. For InternVL2.5-8B, PGDS improves Acc to 74.76% and Target Acc to 50.89%. Notably, PGDS achieves these gains without the heavy computational cost of parameter updates required by LoRA FT. While LoRA FT (Table 8) provides the highest upper bound for classification (e.g., InternVL2.5-8B reaching 83.25% Acc), PGDS offers a cost-effective alternative that bridges the gap between zero-shot and full fine-tuning via policy-driven context engineering.

Model	Chinese Subset (N=200)					English Subset (N=200)				
	Sarcasm		Metaphor		Δ_{F1}	Sarcasm		Metaphor		Δ_{F1}
	Acc.	F1	Acc.	F1		Acc.	F1	Acc.	F1	
GPT-4o	77.00	75.53	41.50	55.51	-20.02	**79.00**	**81.25**	58.00	69.78	-11.47
Gemini-2.5-Flash	77.00	75.27	47.50	57.49	-17.78	72.00	74.31	**64.50**	**72.58**	**-1.73**
Qwen2.5-VL-7B-Instruct	**84.50**	78.77	60.50	59.49	-19.28	67.00	66.67	55.00	66.17	-0.50
InternVL2.5-8B	81.50	**79.10**	**63.75**	**64.02**	-15.08	70.00	63.41	54.00	58.68	-4.73
LLaVA-1.5-7B	35.50	52.40	36.50	53.48	+1.08	57.00	29.51	44.50	17.78	-11.73

Table 9: Binary classification performance (%) comparison between Sarcasm and Metaphor identification across Chinese and English subsets. Δ_{F1} denotes the F1-score drop from sarcasm to metaphor.

Target Recognition Challenges: Despite strong identification results, target recognition remains a bottleneck. Even fine-tuned models (max 53.48% Target Acc) fail to reach the zero-shot performance of Gemini-2.5-Flash (82.54%). This suggests that precise entity localization in sarcasm requires deeper cross-modal reasoning that exceeds the capacity of current small-scale open-source model fine-tuning.

5.3 Probing Deep Semantics: Metaphor vs. Sarcasm

To evaluate the boundaries of deep semantic reasoning, we conduct a comparative study on a subset of 200 high-consistency "Metaphor-Sarcasm" pairs in Chinese and an equal-sized English parallel subset. Results are summarized in Table 9.

Cross-lingual Consistency of Metaphorical Difficulty: Metaphor identification is consistently more challenging than sarcasm detection across languages. In the Chinese subset, the performance gap (Δ_{F1}) for mainstream models (excluding LLaVA) ranges from 15% to 20%. A similar trend is observed in the English subset, where GPT-4o shows an 11.47% drop. This quantifiably validates Gibbs (2017)'s theory that metaphor interpretation necessitates complex cross-domain mapping and deeper cognitive reasoning than sentiment-driven sarcasm.

Cultural Dependency of Model Performance: Model performance is highly correlated with the cultural attributes of their pre-training data. In the **Chinese context**, the open-source InternVL2.5-8B achieves the best metaphor F1 (64.02%), significantly outperforming GPT-4o (55.51%), suggesting that localized multimodal alignment enhances understanding of regional rhetoric. In the **English context**, GPT-4o maintains its lead in sarcasm detection (81.25%), while Gemini-2.5-Flash shows superior adaptability in metaphors with a minimal gap ($\Delta_{F1} = 1.73\%$), indicating robust reasoning in English semantics.

Limitations of Shallow Alignment: LLaVA-1.5-7B fails to demonstrate effective discrimination in both subsets. Its near-zero Δ_{F1} in the Chinese task indicates a reliance on shallow surface features rather than mastering deep semantic alignment, resulting in near-random predictions.

6 Conclusion

We have presented **CFMS**, a fine-grained Chinese multimodal sarcasm benchmark with triple-annotation. To enhance model performance, we propose **PGDS** , a reinforcement learning-based strategy for dynamic demonstration selection, which effectively narrows the gap between zero-shot inference and fine-tuning. Our systematic evaluation reveals that while LoRA-tuned models achieve high classification accuracy, capturing high-order metaphorical logic and assessing AI-generated sarcastic content remain significant challenges. CFMS provides a robust foundation for future research in interpretable affective computing and complex semantic alignment.

Limitations

Despite its contributions, this work has several limitations. First, CFMS is built on Chinese social media data; in the future, we could consider annotating more diverse data to expand its cultural and linguistic scale. Second, while PGDS improves performance, it still faces challenges in high-order metaphorical reasoning and involves higher computational costs than static ICL. Finally, the inherent subjectivity of sarcasm means that individual cognitive biases in interpretation remain unavoidable.

Ethics Statement

This research adheres to the ACL Ethics Policy. All data, comprising Chinese social media samples and English Twitter archives, are sourced from public

platforms. We have strictly anonymized all personal identifiers and ensured that visual content contains no identifiable individuals to protect user privacy. Annotations were performed by students who received fair compensation above local labor standards. We employed a human-in-the-loop approach to verify and correct all LLM-generated pre-annotations, mitigating potential algorithmic biases. The CFMS benchmark and PGDS strategy are intended for beneficial research in explainable AI and creative synthesis; we do not condone their use for generating harmful, deceptive, or discriminatory content.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515.

Elisabeth Camp. 2012. Sarcasm, pretense, and the semantics/pragmatics distinction. *Noûs*, 46(4):587–634.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an obviously perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2025. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Sahar Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Ying Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. *arXiv preprint arXiv:2410.18882*.

Raymond W Gibbs. 2017. *Metaphor wars: Conceptual metaphors in human life*. Cambridge University Press.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Despina Konstantinidou, Dimitrios Karageorgiou, Christos Koutlis, Olga Papadopoulou, Emmanouil Schinas, and Symeon Papadopoulos. 2025. Navigating the challenges of AI-generated image detection in the wild: What truly matters? *arXiv preprint arXiv:2507.10236*.

Bin Liang, Chenwuei Lou, Xin Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. 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)*, pages 1767–1777.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Zeyu Lu, Di Huang, Lei Bai, Jingjing Qu, Chengyue Wu, Xiaoxian Liu, and Wanli Ouyang. 2023. Seeing is not always believing: Benchmarking human and model perception of AI-generated images. In *Advances in Neural Information Processing Systems*, volume 36, pages 25435–25447.

Robert Munro. 2021. *Human-in-the-loop machine learning: Active learning and annotation for human-centered AI*. Manning Publications.

Jonas Oppenlaender. 2024. A taxonomy of prompt modifiers for text-to-image generation. *Behaviour & Information Technology*, 43(15):3763–3776.

Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. MMSD2.0: Towards a reliable multi-modal sarcasm detection system. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10834–10845.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models

from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Rossano Schifanella, Paloma De Juan, Joel R Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM International Conference on Multimedia*.

Margot Van Mulken, Renske Van Enschot-van Dijk, and Hans Hoeken. 2005. Puns, relevance and appreciation in advertisements. *Journal of pragmatics*, 37(5):707–721.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval*, pages 641–649.

Tianqi Yue, Xiaochen Shi, Rui Mao, Zheheng Hu, and Erik Cambria. 2024. SarcNet: A multilingual multimodal sarcasm detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14325–14335.

A Appendix

A.1 AI-Generated Sarcastic Content

Figure 5 illustrates sarcastic images generated by different MLLMs. We observe that while models can generate visual associations, they struggle to construct deep "sarcastic logic" involving semantic conflicts between text and visual features.

Figure 5: Examples of AI-generated images under sarcastic prompts.

A.2 Prompt Templates

A.2.1 GPT-4o Pre-annotation Prompt

The following prompt employs a Chain-of-Thought (CoT) and critical thinking framework to assist in high-quality initial labeling.

Original Chinese Prompt:
请严格遵循批判性思维原则分析<文本-图片对>，按步骤评估讽刺可能性：
图片所配文本为：{文本} (可能为空)
【分析框架】
1. 显性要素记录：文本字面含义；图片客观描述；文化语境标注。
2. 矛盾检测（需满足至少两项）：图文语义冲突；情感基调错位；符号系统反常；存在双层含义线索。
3. 反假设验证：替代解释测试；作者意图考量；受众认知调查。
【输出规范】
确认讽刺存在的必要条件：①存在可验证的语义对立；②符合常规讽刺表达范式；③排除字面解释合理性。
否定讽刺的充分条件：①图文存在自洽的非讽刺解释；②矛盾强度低于文化认知阈值；③缺乏双层意义证据。
注意：不要把好的现象强行解释为讽刺，不确定时统一判为无讽刺。输出对象需简短。
【结构化输出】
<result> <讽刺对象>...</讽刺对象> <讽刺解释>...</讽刺解释> </result>

English Translation:
Please strictly follow critical thinking principles to analyze the <image-text pair> and evaluate the probability of sarcasm step-by-step:
The caption provided is: {text} (may be empty)
[Analysis Framework]
*1. **Explicit Element Recording**: Literal meaning of text; objective description of image; cultural context tagging.*
*2. **Contradiction Detection** (At least two required): Semantic conflict; emotional misalignment; anomalous use of symbols; clues of double meaning.*
*3. **Counter-hypothesis Validation**: Alternative explanation testing; author intent; audience perception survey.*
[Output Norms]
Necessary conditions: (1) Verifiable semantic opposition; (2) Alignment with sarcastic paradigms; (3) Exclusion of literal plausibility.
Sufficient conditions to negate: (1) Self-consistent non-sarcastic interpretation; (2) Contradiction below cognitive thresholds; (3) Lack of double meaning evidence.
Note: Do not interpret positive phenomena as sarcasm. If uncertain, label as non-sarcastic.
[Structured Output]
<result> <target>...</target> <explanation>...</explanation> </result>

A.2.2 Evaluation and Generation Prompts

Table 10 provides the prompts for zero-shot detection, few-shot ICL, and AI-guided generation.

Task	Original Chinese Prompt	English Translation
Zero-shot	给你一张图片，图片配文为:{text}。分析该图文对是否含有讽刺，并给出讽刺对象和解释...	Given an image with caption: {text}. Analyze whether this pair contains sarcasm, and provide the target and explanation...
Few-shot	Example: 输入：配文:{text}; 图片:{img}. 输出：是否讽刺:{y/n}; 讽刺对象:{target}; 讽刺解释:{exp}. Test: 给你一张图片,分析该图片是否含有讽刺...	Example: Input: {text/img}; Output: {y/n; target; exp}. Test: Given an image, analyze whether it contains sarcasm and provide the target/explanation.
AI Gen	请生成一幅讽刺图片，图片描述是：{描述}；讽刺对象为：{对象}；讽刺解释为：{解释}。	Please generate a sarcastic image. Description: {desc}; Target: {target}; Explanation: {exp}.

Table 10: Standardized prompts for model evaluation and generation tasks.

A.3 Annotation Interface

Figure 6 displays the custom Web-based annotation platform, which supports synchronized image-text viewing, target selection, and multi-stage verification.

Figure 6: Snapshot of the custom-developed CFMS annotation platform.