Reading the Lines, Decoding the Minds: Explainable Satirical Cartoon Detection via On-Canvas and Beyond-Canvas Modeling

Anonymous ACL submission

Abstract

Multimodal Satire Detection (MSD) aims to identify implicit criticism and stance in social media. However, existing MSD benchmarks mainly use user generated content and predominantly consist of realistic photographs, where satire is often conveyed through overt image text incongruity. By contrast, increasingly popular editorial cartoons convey satire through symbolic and metaphorical cues, requiring models to infer latent intent and pragmatics. To address these limitations, we propose MSCE, a dedicated evaluation benchmark for Multimodal Satirical Cartoons Evaluation. As a strong baseline for MSCE, we introduce a novel explainable framework named DOUBLE. DOUBLE endows Multimodal Large Language Model (MLLM) with a Clue2View mechanism to explicitly model interpretation through two complementary views: a literal on-canvas view and a metaphorical beyond-canvas view. In addition, DOUBLE incorporates a lightweight Small Language Model Arbiter to distill reasoning traces from the MLLM, ensuring reliable predictions with lower computational costs. Experiments on MSCE demonstrate that DOUBLE achieves the best performance, showcasing its ability to provide clear, well-grounded rationales for complex satirical cartoons.

1 Introduction

Satire has become a prevalent rhetorical device on social media, enabling users to comment on public affairs and signal stance. Satirical posts often employ overtly incongruent expressions to implicitly criticize or attack a target (Farabi et al., 2024; Tiwari et al., 2023). Effective identification and interpretation of multimodal satire is therefore consequential for a range of downstream applications, including misinformation detection (Rubin et al., 2016; Bedard and Schoenthaler, 2018; De Sarkar et al., 2018), public opinion monitoring (Lu et al., 2025), and sentiment analysis (Joshi et al., 2017; Mishra et al., 2016).

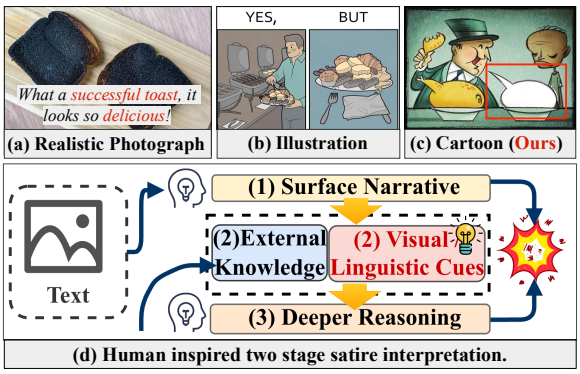

Figure 1: Comparison of image types in multimodal satire detection datasets. & Human inspired two stage satire interpretation.

Recently, the community has established several multimodal satire benchmarks on open social platforms, such as MMSD (Cai et al., 2019), MMSD2.0 (Qin et al., 2023), and SarcNet (Yue et al., 2024). They also have facilitated standardized evaluation and rapid methodological progress. A common assumption in prior work is that cross modal incongruity is a primary cue for satire (Pan et al., 2020; Liang et al., 2021). Models can leverage this incongruity to detect satire and achieve state of the art performance in open domain social media (Liang et al., 2022; Qiao et al., 2023; Wang et al., 2024a).

However, when the scope of investigation extends to satirical cartoons, existing benchmarks, which primarily feature realistic photographs or illustrations, exhibit notable limitations in two respects. (1) Most existing datasets are simply and directly collected from social media user generated content, where creators often elicit satire via explicit image text conflicts. As illustrated in Figure 1(a) and (b), the satire is easy to infer because it hinges on an explicit mismatch, either the caption praising "delicious" food despite burnt toast or the stark contrast between excessive stockpiling and subsequent waste. However, when satire is conveyed primarily through such obvious surface-level

cues, models may shortcut the reasoning process and simply learn to detect salient anomalies. (2) Recently, cartoons are a widely used for satire in news commentary and public discourse (Kundu et al., 2025). Crucially, cartoon satire rarely relies on overt cues. Instead, it emerges from the combination of symbolism and metaphor, suggestive narratives, background knowledge about events and eras, and intent inference. As a result, when satirical cues are veiled and the cartoon conveys an underlying critical stance, models are more prone to misclassification. For example, Figure 1(c) conveys satire through symbols (top hat & suit $\rightarrow$ capitalist; frail body $\rightarrow$ working class), metaphor (devouring $\rightarrow$ predatory exploitation), and context (empty speech bubble $\rightarrow$ hollow promises). Therefore, existing Multimodal Satire Detection (MSD) benchmarks may fall short in evaluating non contradiction based satire, motivating a dedicated benchmark for editorial cartoons.

To address these limitations, we introduce **MSCE**, the first Multimodal Satirical Cartoon Evaluation benchmark, which contains 2,016 image–text pairs collected from satirical cartoons published in the *Satire and Humor* supplement of *People's Daily*. MSCE focuses on cartoons whose image and text are surface consistent, where satire arises from implicit intent rather than shallow conflicts. Furthermore, it evaluates core abilities for MSD, including symbol recognition, metaphor mapping, contextual and era specific knowledge use, and intent conflict inference. However, most existing satire detection approaches treat cross modal incongruity as the primary cue, and they encounter two major challenges on MSCE as follows:

C1: From surface consistency to metaphorical reasoning. Unlike conventional MSD, where satire is typically signaled by explicit cross modal incongruity, editorial cartoons often maintain surface consistency between text and image. And the surface consistency conceals the underlying satirical intent. Detecting such satire therefore requires metaphorical understanding beyond surface semantics. In particular, models must perform multi-hop reasoning that maps concrete visual symbols, such as a broken bowl, to abstract concepts like economic crisis, a capability that contradiction-based approaches fail to capture.

C2: Complexity of disentangling Literal Narrative and Satirical Intent. In editorial cartoons, satirical cues are interwoven with literal narratives, so indiscriminate fusion can let salient but non decisive signals dominate and increase misclassification. Even MLLMs may resort to coherent post hoc stories without isolating true satirical triggers. Therefore, effective satire detection requires explicit multi-perspective modeling: one encodes the on-canvas literal narrative, another captures beyond-canvas contextual intent, and the final prediction is derived from their contrast.

To address the challenges, we propose DOUBLE, a novel cartoon-oriented satire detection framework powered by **D**eriving **O**n-canvas **U**nderstanding and **B**eyond-canvas **L**ogic **E**xtraction. DOUBLE can explicitly models the reasoning process to yield clear, evidence-grounded rationales. It is inspired by the progressive way humans interpret satirical cartoons (see Figure 1 (d)). (1) Readers first form a surface narrative from on-canvas content, and (2) then use salient cues to retrieve beyond-canvas context to (3) infer a deeper target and intent. Consequently, satire arises from the contrast between the surface reading and the context-enriched interpretation. Accordingly, we introduce **Clue2View**, a dual-view mechanism that generates: (i) a literal *non-satirical* explanation grounded in on-canvas cues, and (ii) a context-enriched *satirical* explanation by integrating beyond-canvas knowledge, enabling models to move beyond surface consistency and better detect metaphor-driven satire.

Although fine-tuning multimodal large language models (MLLMs) with Clue2View substantially improves performance on MSD, it incurs expensive computational cost and limits practical usability. To address this issue, we further propose a Small Language Model Arbiter (**SLM Arbiter**) that serves as an efficient and reliable decision maker within DOUBLE. Built upon a small language model, SLM Arbiter distills knowledge from Clue2View to refine the outputs of MLLMs through lightweight fine-tuning, ensuring reliable predictions with low computational cost. We summarize our main contributions as follows.

- We further introduce MSCE, the first benchmark for satirical cartoon detection that focuses on veiled multimodal satire and requires deep contextual and symbolic reasoning.

- We propose DOUBLE, an explainable framework for MSD that is inspired by this progressive human reasoning process. As a strong

baseline on MSCE, DOUBLE enables reliable detection even when image and text are surface consistent, while providing verifiable evidence.

- Extensive experiments and case studies on MSCE show that DOUBLE substantially outperforms existing approaches for MSD, and it establishes a strong baseline for future research.

The code and data for our proposed DOUBLE and MSCE are available at https://anonymous.4open.science/r/DOUBLE-5577/.

2 Related Work

2.1 Multimodal Satire Detection Benchmarks

Various benchmarks have been introduced for MSD. Cai et al. (Cai et al., 2019) formulated MSD on Twitter image–text posts and released the MMSD dataset. Qin et al. (Qin et al., 2023) subsequently refined MMSD by removing spurious cues and proposed MMSD2.0. Further, Desai et al. (Desai et al., 2022) introduced the task of multimodal satire explanation and constructed the MORE dataset, where each instance is equipped with natural language explanation. To evaluate the capabilities of vision language models, the YesBut benchmark adopted a multi task setup to assess whether models truly understand and can explain satire (Nandy et al., 2024). In the cross lingual setting, SarcNet (Yue et al., 2024) provides English and Chinese image–text pairs, complementing multilingual evaluation for MSD.

Most existing datasets rely primarily on realistic photographic images, which tends to simplify satire understanding. Even YesBut, despite including cartoonist produced illustrations, offers limited coverage of core cartoon satire mechanisms, including symbolism, metaphor, and contextual inference. To address these gaps, we introduce a benchmark of multimodal satirical cartoons that emphasizes implicit satire and demands complex contextual reasoning, aiming at promoting context driven satire understanding.

2.2 Multimodal Satire Detection Methods

Most existing approaches to MSD are grounded in modality incongruity, assuming that satire arises from semantic or affective conflict between text and vision (Cai et al., 2019; Pan et al., 2020; Liang et al., 2022). Despite strong performance on social media benchmarks, these methods are limited in visual symbol grounding and deep metaphor reasoning, both of which are crucial for satirical cartoons and are often cross modal consistent.

With the rapid progress of multimodal foundation models, recent work leverages LLMs to reformulate MSD. (Tang et al., 2024) uses in-context learning with retrieved similar demonstrations, (Jana et al., 2024) applies soft prompt tuning to highlight cross modal conflicts, and (Wang et al., 2024b) employs a multiagent pipeline that combines semantic and affective analyses for final prediction. However, these methods have two key limitations. First, they rarely provide verifiable evidence and explicit reasoning, making errors hard to diagnose and explanations less reliable. Second, they inadequately model visual symbols and metaphors.

To address this limitation, we propose novel DOUBLE, which consists of Clue2View and SLM Arbiter. It detects satire by measuring the intentional divergence between two views, while explicitly modeling the underlying reasoning process, which remains informative even under surface cross modal consistency. DOUBLE further employs a lightweight SLM Arbiter to refine MLLM outputs, mitigating hallucinations while maintaining effective and efficient inference.

3 MSCE Benchmark

In this study, we introduce MSCE, a benchmark of editorial cartoons where satire is conveyed through symbolism, metaphor, and contextual inference rather than explicit image-text contradiction.

3.1 Data Collection

We construct our benchmark by crawling editorial cartoons from *Satire and Humor* , a comic supplement of *People's Daily*. The collected cartoons span a diverse set of topics, including international affairs, domestic social news, and issues related to everyday livelihood. In total, we gather 2,016 cartoon instances. Each instance contains a title and the cartoon image, and a subset further includes editor-provided background descriptions released by *Satire and Humor*.

For samples whose background descriptions are excessively long, we manually rewrite the descriptions to improve conciseness and consistency, ensuring that each rewritten background context is limited to approximately 100 Chinese characters

Text Input	Class	Train	Validation	Test	Total
Title	Sarcastic Non-Sarcastic	801 (73.49%) 289 (26.51%)	44 (70.97%) 18 (29.03%)	91 (71.65%) 36 (28.35%)	1,279 (63.44%)
Title + Background	Sarcastic Non-Sarcastic	419 (67.26%) 204 (32.74%)	25 (65.79%) 13 (34.21%)	55 (72.37%) 21 (27.63%)	737 (36.56%)
Total	Sarcastic Non-Sarcastic	1,220 (71.22%) 493 (28.78%)	69 (69.00%) 31 (31.00%)	146 (71.92%) 57 (28.08%)	2,016

Table 1: Distribution of satiric and non-satiric samples across dataset splits under different text-input settings.

while preserving the essential context required to interpret the satire.

3.2 Dataset Annotation

To ensure high-quality annotations, we conduct manual labeling using the open-source platform Label Studio (Tkachenko et al., 2020-2025). Three master's students conducted three rounds of independent annotation and verification. All annotators are native Chinese speakers and possess sufficient linguistic proficiency and reading comprehension to reliably identify satirical intent in cartoons.

Prior to annotation, we randomized the distribution of samples to ensure each annotator received a diverse subset of cartoons. During the labeling process, annotators were presented with either (i) the cartoon image and its title, or (ii) the image, title, and accompanying background context. Ground truth labels were determined via majority vote. Detailed annotation instructions can be found in Appendix A. We assess inter-annotator reliability using Fleiss' Kappa and obtain a score of 0.6302, indicating substantial agreement among annotators (Landis and Koch, 1977).

3.3 Dataset Analysis

Table 1 reports the class distribution across all splits. Overall, (1) the benchmark contains 2,016 instances, with 1,713/100/203 samples in train/validation/test, respectively. (2) Most samples provide only the title as textual input (63.44%). (3) Sarcastic cartoons dominate all splits (69%–72%), indicating a stable class distribution.

Figure 2 reports topic-level statistics of our dataset. Figure 2 (a) shows the relative prevalence of 11 topics, with Economy accounting for the largest share (15.8%), followed by Workplace (14.5%) and Education (14.1%). Technology (12.6%) and Environment (11.5%) also constitute substantial portions, whereas Real Estate is the least represented topic (0.7%). Figure 2 (b)

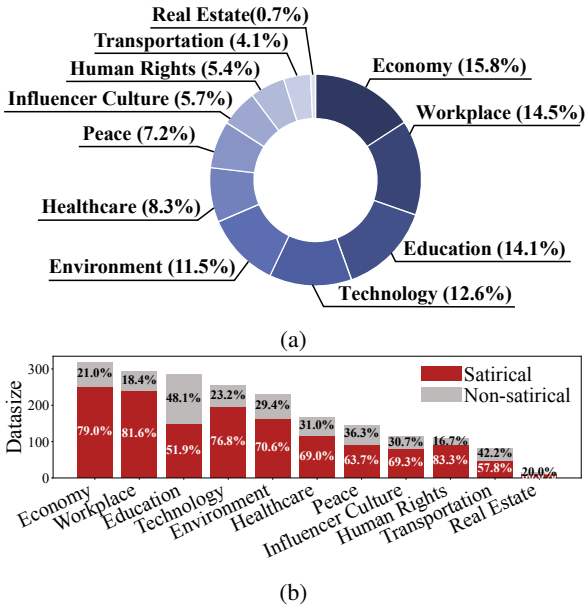

Figure 2: Dataset statistics by topic: (a) topic distribution and (b) satire distribution within each topic.

further breaks down each topic into sarcastic and non-sarcastic subsets. Satire is dominant in most topics, while Education is nearly balanced between the two classes.

4 Methodology: DOUBLE

Problem Formulation. Let $\mathcal{C} = \{\mathcal{T}, \mathcal{V}\}$ denote an editorial cartoon instance, comprising a textual modality $\mathcal{T}$ and a visual modality $\mathcal{V}$. The goal of MSD is to determine whether $\mathcal{C}$ conveys *satire* by jointly modeling both modalities.

Overview. We first propose Clue2View, a dual-view explanation mechanism for MSD. It explicitly models interpretation through two complementary views. (1) On-Canvas View guides MLLMs to parse the cartoon and text into a Literal Scene Graph (LSG), which captures explicitly depicted entities and their spatial relations, providing a faithful description of the visual content. Based on the LSG, the model then produces a *non-satirical*

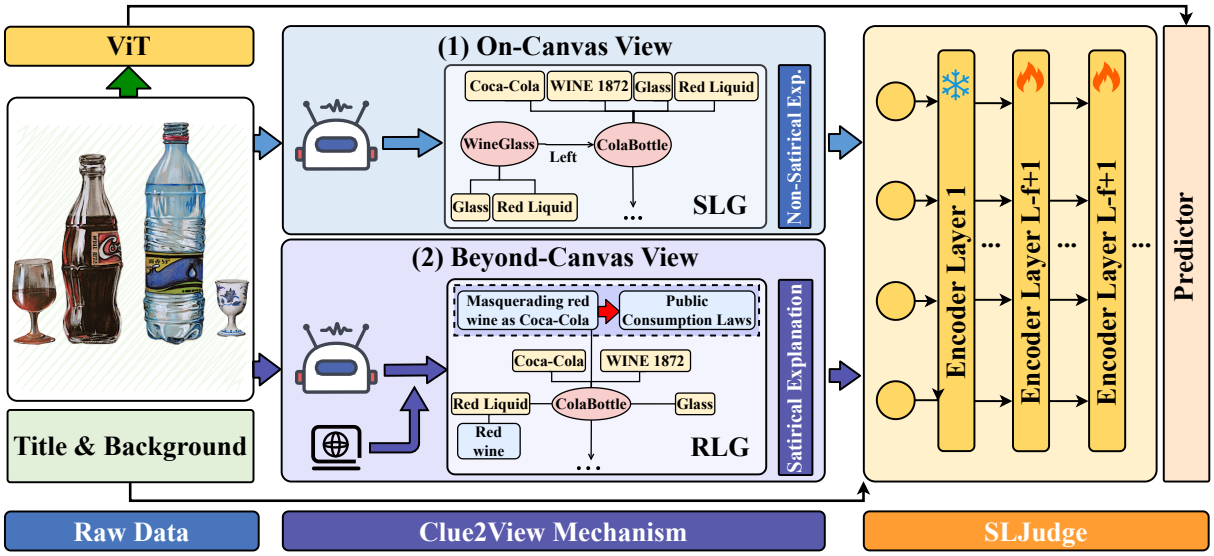

Figure 3: The structure of our proposed DOUBLE framework. (1) Clue2View provides complementary literal and satirical interpretations based on the Literal and Refined Scene Graphs, respectively. (2) The SLM Arbiter is to distill the reasoning traces from Clue2View to facilitate reliable satire detection.

interpretation that summarizes the literal, factual content. (2) Beyond-Canvas View further instructs MLLMs to enrich the LSG with commonsense and world knowledge, yielding a Refined Scene Graph (RSG) that explicitly represents intent conflicts underlying satire. Conditioned on the RSG, the model generates a *satirical* interpretation that explains the implied meaning. We further introduce SLM Arbiter to distill and validate the reasoning traces from MLLMs, improving the reliability of the final prediction. Figure 3 provides an overview of our proposed DOUBLE framework.

4.1 Clue2View Mechanism

For the textual modality $\mathcal{T}$, we extract the cartoon metadata, including the title and the background context (when available), and denote it as $\mathcal{T}_m$. For the visual modality, we encode the cartoon image and obtain its visual representation, denoted as $\mathcal{I}_c$. The detailed feature extraction procedure is provided in Appendix D.

4.1.1 On-Canvas View

Cartoonists often conceal their critical stance behind seemingly ordinary scenes, intentionally creating a gap between the literal depiction and the underlying intention. Existing multimodal models often rely primarily on shallow text image alignment cues, rather than high level evidence essential for satire such as intent conflicts and implied stances. Consequently, it is necessary to reorganize raw inputs into structured, reasoning-oriented clues. To address these problems, Clue2View first refines the multimodal content by extracting salient visual and textual clues to construct a multimodal scene graph, and then rewrites the graph into semantically grounded statements.

We first develop a vision-centric strategy to refine the literal scene content of multimodal comics. The strategy retains only information that is directly observable in the image and the accompanying text. Specifically, it preserves explicit entities, relations, and attributes while discarding external world knowledge, cultural assumptions and information that requires metaphorical interpretation. This restriction yields an objective and intuitive representation that is grounded in the comic itself.

Given a multimodal comic instance $\mathcal{C}$, we generate a literal scene graph $\mathcal{L}_{sg}$ following the template prompt Prompt_1 in Step 1 of Figure 6 in Appendix D. The generation process is:

$$\mathcal{L}_{sg} = \mathcal{F}(\mathcal{C}, \text{Prompt}_1), \tag{1}$$

where $\mathcal{F}(\cdot)$ denotes the MLLM.

Moreover, since a non-sarcastic reading is essentially an objective statement of the explicit content conveyed by the sample, we further derive a non-sarcastic interpretation from the $\mathcal{L}_{sg}$. Specifically, guided by the template prompt Prompt_2 in Step 2 of Figure 6 in Appendix D, we obtain the non-sarcastic interpretation: $\mathcal{V}_{non}$, defined as:

$$\mathcal{V}_{non} = \mathcal{F}(\mathcal{L}_{sg}, \text{Prompt}_2). \tag{2}$$

4.1.2 Beyond-Canvas Sarcastic Interpretation

This step guides the MLLMs to construct a refined scene graph $\mathcal{R}_{sg}$ for a given multimodal satirical comic. We construct $\mathcal{R}_{\text{sg}}$ from the $\mathcal{L}_{\text{sg}}$ and a fact-oriented description $\mathcal{D}$. The description serves as a high-level semantic anchor, which specifies the retrieval intent and constrains the integration of relevant world knowledge. In other words, this design reduces semantic drift during knowledge injection and emphasizes the satirical focus. As a result, the $\mathcal{R}_{sg}$ integrates external commonsense knowledge with the information in D, which improves structured modeling of satirical semantics. For a multimodal comic sample $\mathcal{C}$, the $\mathcal{R}_{sg}$ is generated using the template Prompt_3 in Step 3 of Figure 6 in Appendix D. We formalize this process as:

$$\mathcal{R}_{sg} = \mathcal{F}\big([\mathcal{L}_{sg}; \mathcal{D}], \text{Prompt}_3\big), \tag{3}$$

where $[\,;\,]$ denotes concatenation.

Based on the semantic relations and metaphorical cues provided by the $\mathcal{L}_{sg}$ and the $\mathcal{R}_{sg}$, the model can further perform a deeper sarcastic interpretation and inference. The model identifies key semantic conflicts and forms explicit reasoning traces. These traces explain the satirical interpretation, thereby enhancing both interpretability and transparency. The sarcastic reasoning output is guided by the template Prompt_4 in Step 4 of Figure 6 in Appendix D, and is formalized as:

$$\mathcal{V}_{satire} = \mathcal{F}([\mathcal{L}_{sg}; \mathcal{R}_{sg}], \text{Prompt}_4). \tag{4}$$

By leveraging the logically rigorous and coherent Clue2View mechanism, our framework effectively addresses key limitations of prior work (Wang et al., 2024b; Tang et al., 2024; Jana et al., 2024). Existing MSD methods often rely primarily on cross-modal incongruity, overlook hallucinations in MLLMs predictions, and provide limited semantic analysis for metaphor-rich cartoons, which leads to weak performance on non-obvious satire. In contrast, Clue2View explicitly models both literal semantics and deeper intentions, and derives complementary reasoning traces from these representations. This explicit reasoning modeling yields more informative evidence for decision making and translates into substantial gains in prediction accuracy.

4.2 Small Language Model Arbiter

Although MLLMs can produce predictions through our carefully designed four-step pipeline, satire detection fundamentally relies on holistic and intuitive cognition rather than a sequence of logical inferences (Zhang et al., 2025). This characteristic makes it difficult for MLLMs to reach stable and accurate judgments. Moreover, MLLMs suffer from systematic unfaithfulness (Turpin et al., 2023): their intermediate reasoning and generated rationales may not faithfully reflect the actual decision process, and can even contain hallucination. While task-specific fine-tuning of MLLMs for MSD can partially mitigate these issues, the practical utility of this approach is limited by the massive parameter count of the MLLMs and the corresponding computational costs.

To address these limitations, we propose SLM Arbiter, which distills the outputs of MLLM into a substantially SLM. By fine-tuning this compact model, we obtain a more reliable and deployable solution with far fewer trainable parameters and markedly improved computational efficiency. The fine-tuning procedure can be formulated as:

$$\mathcal{S}(x) = \mathcal{L}_L \circ \mathcal{L}_{L-1} \circ \cdots \circ \mathcal{L}_{L-f+1} \circ \mathcal{L}^{\text{fixed}}_{L-f} \circ \cdots \circ \mathcal{L}^{\text{fixed}}_1(x), \tag{5}$$

where $\mathcal{L}_i$ denotes the i-th Transformer encoder layer. During training, only the final f layers, from $\mathcal{L}_{L-f+1}$ to $\mathcal{L}_L$, are updated, while all earlier layers remain frozen. This design preserves the capacity of SLM Arbiter to audit and verify the outputs of the MLLM, while achieving significantly lower training and inference costs.

We concatenate the text $\mathcal{T}_m$ from a satirical cartoon with a non-satirical interpretation $\mathcal{V}_{non}$ and a satirical interpretation $\mathcal{V}_{satire}$, and feed the resulting sequence into SLM Arbiter to obtain the textual feature representation required for prediction:

$$\mathbf{H}_t = \mathcal{S}\big([\mathcal{T}_m;\, \mathcal{V}_{non};\, \mathcal{V}_{satire}]\big), \tag{6}$$

where n denotes the concatenated sequence length and d_t is the hidden dimension. For the visual representation $\mathbf{H}_v \in \mathbb{R}^{n_v \times d_v}$, we extract features from the cartoon image $\mathcal{I}_c$ using a pretrained Vision Transformer (ViT) (Zhang and Yang, 2021): $\mathbf{H}_v = \text{ViT}(\mathcal{I}_c)$, where n_v and d_v denote the numbers of patches and the feature dimension, respectively.

To align the text feature $\mathbf{H}_t$ and the visual feature $\mathbf{H}_v$, we use MLP-based nonlinear projections $\Psi_t(\cdot)$ and $\Psi_v(\cdot)$ to map them into a shared space. We fuse the projected features with an attention-weighted sum and feed the fused vector to an MLP predictor to obtain $\hat{y} = \text{Predictor}(\text{Fusion}(\Psi_t(\mathbf{H}_t), \Psi_v(\mathbf{H}_v)))$ During training, we optimize the parameters of SLM

Methods	ACC	M-F1	M-P	M-R
Docmsu	72.41	65.85	65.85	65.85
DAIE	81.77	72.94	**82.42**	70.22
MMSD	67.49	60.96	60.71	61.36
MMSD2.0	76.85	68.17	71.43	66.79
S^3 Agent+GPT-4o-mini	71.92	67.62	66.98	69.25
S^3 Agent+GPT-5.1	82.27	76.12	79.39	74.30
S^3 Agent+Qwen2.5-VL	73.89	65.95	67.08	65.27
COC+GPT-4o-mini	66.90	65.92	68.96	73.31
COC+GPT-5.1	70.44	69.31	71.69	76.78
COC+Qwen2.5-VL	78.82	75.15	74.22	76.72
GOC+GPT-4o-mini	74.88	66.39	68.34	65.42
GOC+GPT-5.1	79.31	75.61	74.71	77.06
GOC+Qwen2.5-VL	76.35	61.15	75.16	60.57
GPT-4o-mini	80.30	74.47	76.22	72.40
GPT-5.1	75.86	72.16	71.29	74.13
Qwen2.5-VL	81.28	77.07	76.80	77.36
DOUBLE+GPT-4o-mini	**85.22**	81.90	81.58	82.24
DOUBLE+GPT-5.1	84.73	81.57	80.87	82.43
DOUBLE+Qwen2.5-VL	**85.22**	**82.26**	81.43	**83.31**

Table 2: Performance comparison on the MSCE dataset. The best results are highlighted in bold, while the second-best results are indicated with underlines. Higher values of ACC, M-F1, M-P, and M-R indicate better performance.

Arbiter by minimizing the binary cross-entropy loss.

5 Experiments

5.1 Experimental Setup

In this section, we present a simple experimental setup, and the detailed version can be found in the Appendix C.

Baselines. To assess the effectiveness of DOUBLE, we compare it with several competitive baselines in MSD, including SarcasmCue (Yao et al., 2025), S^3 Agent (Wang et al., 2024b); Docmsu (Du et al., 2024), DAIE (Wu et al., 2025), MMSD (Cai et al., 2019), MMSD2.0 (Qin et al., 2023).

Metrics. Following prior works (Yao et al., 2025; Hong et al., 2025), we evaluate all models on our proposed MSCE benchmark and employ four metrics : Accuracy (**ACC**), Macro-F1 scores (**M-F1**), Macro Precision (**M-P**) and Macro Recall (**M-R**).

5.2 Main Performance

To assess the effectiveness of DOUBLE, we compare it with 10 competitive baselines, with results reported in Table 2. We draw four key observations.

First, DOUBLE achieves the best overall performance on MSCE across nearly all metrics, improving Accuracy by 12.62% and Macro F1 by 18.06% over the average baseline. The advantage remains consistent when switching backbones, indicating robust gains rather than backbone dependent tuning.

Second, methods relying on shallow cross modal cues are insufficient for handling the deep metaphorical expressions that frequently appear in MSCE. Such approaches typically emphasize surface level cross modal inconsistency. However, the sarcastic intent is often implicit in MSD, the metaphor is deeper, and cross modal conflict may not be explicitly manifested. As a result, these models struggle to capture the core semantics, which fundamentally limits their overall performance.

Third, longer and more complex reasoning chains do not necessarily help. Multiagent reasoning and chain based approaches such as S^3 Agent, COC, and GOC show unstable gains and even notable degradation on some backbones, suggesting that extended reasoning is vulnerable to hallucination and noise accumulation that corrupts crucial evidence.

Finally, while MLLMs based methods are strong in zero-shot multimodal settings, they are less reliable on MSD without task specific adaptation, and full fine tuning is often prohibitively expensive. By refining raw the outputs of MLLMs through SLM Arbiter, DOUBLE provides more stable predictions and achieves better MSD performance than conventional MLLMs based approaches.

5.3 Ablation Study

Effect of Clue2View Mechanism. To quantify the contribution of each component in Clue2View mechanism, we conduct an ablation study that isolates individual steps and measures their impact on overall performance. Specifically, we evaluate four variants: (1) *w/o LSG*, which removes the LSG extraction module; (2) *w/o RSG*, which removes the RSG extraction module; (3) *w/o SG*, which removes both LSG and RSG and disables the dual view interpretation mechanism accordingly; and (4) *w/o Dual View*, which removes dual view interpretation and models the task using only scene graph description.

The results indicate that each step of Clue2View is essential for satirical cartoon detection. Beyond improving accuracy, these components also provide useful interpretability by offering reasoning evidence that supports the final prediction. Notably, removing the RSG mechanism yields the largest performance drop. This suggests that satirical cartoons often convey intent through implicit mean-

Module	MLLM Backbone	Variant	ACC	M-F1
Clue2View	GPT-4o-mini	w/o LSG	84.73	80.33
		w/o RSG	82.76	78.54
		w/o SG	84.24	79.05
		w/o Dual-View	84.73	81.57
	GPT-5.1	w/o LSG	84.24	79.57
		w/o RSG	83.74	79.54
		w/o SG	84.24	79.05
		w/o Dual-View	84.24	80.89
	Qwen2.5-VL	w/o LSG	83.74	79.06
		w/o RSG	83.25	79.04
		w/o SG	84.24	79.05
		w/o Dual-View	85.22	81.71
SLM Arbiter	GPT-4o-mini	w/o Fine-tune	79.31	73.81
	GPT-5.1	w/o Fine-tune	79.31	73.51
	Qwen2.5-VL	w/o Fine-tune	78.82	73.04
DOUBLE	GPT-4o-mini	ALL	85.22	81.90
	GPT-5.1	ALL	84.73	81.57
	Qwen2.5-VL	ALL	85.22	82.26

Table 3: Ablation study on key components of DOUBLE.

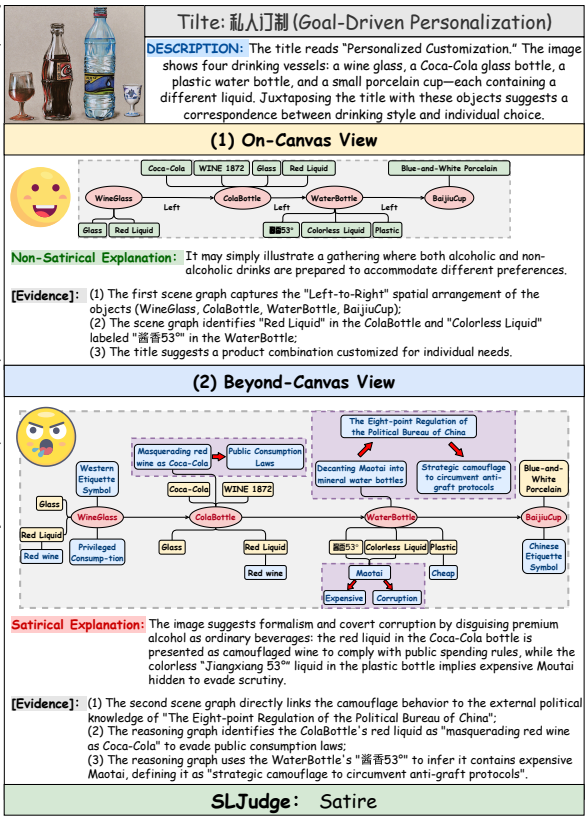

Figure 4: Case study of correctly detected comic satire on the MSCE dataset.

ings and deeper metaphors, where relying solely on surface level clues is insufficient to consistently capture the key semantics. In contrast, the deep semantic reasoning cues provided by RSG are more critical for robust discrimination.

Effect of SLM Arbiter. To assess the effectiveness of SLM Arbiter, we introduce an ablation variant *w/o Fine tune*, where the SLM is not fine tuned. The results show that fine tuning SLM Arbiter is necessary to adapt to the distilled outputs produced by the MLLM. Freezing all layers leads to a substantial performance degradation, confirming the importance of parameter updating for effective alignment with the distillation signals.

5.4 Case Study

To further examine the interpretability of our proposed DOUBLE, we randomly sample a cartoon from the MSCE dataset and analyze how DOUBLE infers its satirical label, as shown in Figure 4. The cartoon depicts several drink containers and each associated with different beverages. DOUBLE first constructs a LSG and produces an intuitive non-satirical reading, namely that the image portrays a gathering that prepares both alcoholic and non alcoholic drinks to accommodate different preferences. It then incorporates relevant commonsense and contextual knowledge, such as the Central Eight Point Regulations and related laws, to refine the scene representation. Based on the resulting RSG, DOUBLE performs further reasoning and derives the underlying satirical interpretation, which uses the surface narrative of 'private customization' to criticize deliberate rule breaking and concealed corruption.

This case illustrates that DOUBLE not only detects satirical content, but also provides a coherent and well supported justification grounded in commonsense knowledge and logical inference.

6 Conclusion

This work introduces MSCE, the first official media sourced benchmark for MSD. MSCE emphasizes hard cases where image text pairs appear coherent, but satire arises from metaphor, symbolism, and culture specific context, demanding nontrivial cross modal reasoning. We also propose DOUBLE, an explainable framework for MSD. Through Clue2View, DOUBLE progressively contrasts a literal reading with the implied intent. A Small Language Model Arbiter further distills the dual view traces from the MLLM to support efficient and reliable decisions, reducing cost and filtering hallucinated evidence. Experiments and case studies show that DOUBLE achieves strong performance with clear rationales, providing a competitive baseline for MSD.

7 Limitations

Cross-cultural and time-sensitive generalization. Satire interpretation is strongly conditioned on cultural context and current affairs. Even with Beyond-Canvas knowledge enrichment, cartoons from different regions and languages can differ substantially in symbolism, sociopolitical framing, metaphor conventions, and textual expression, which may reduce robustness under domain shift. Moreover, DOUBLE relies on the coverage and freshness of the underlying MLLMs knowledge; for newly emerging events or memes, the model may lag or misinterpret. Biases in long-tail or niche knowledge can further affect the construction of the refined scene graph and the induced intent conflicts.

Subjectivity and annotator agreement. Satire is inherently subjective, and judgments vary with annotators' background knowledge, cultural assumptions, and personal stance. Despite using standardized guidelines and consistency checks, disagreements and borderline cases are difficult to eliminate. This residual label noise can blur decision boundaries and make evaluation sensitive to small changes in annotation criteria.

8 Ethical considerations

We have obtained the necessary authorization to collect and use content from People's Daily Online, as the website's publicly available terms grant rights to users who comply with its online rules. We carefully reviewed these terms and adhered strictly to all stated requirements throughout data acquisition and annotation. In the revised manuscript, we provide a detailed account of the data sources, the collection pipeline, and the annotation procedure to ensure transparency and reproducibility. We further commit to continuous oversight of the dataset and will promptly remove or revise any data that could be considered ethically inappropriate.

This work advances multimodal sarcasm detection to enable responsible research use and to deepen our understanding of sarcastic expression in real-world communication, thereby supporting content analysis and related downstream applications. We acknowledge that techniques developed for sarcasm detection may be misused to facilitate the creation of misleading content that is more difficult to identify. We explicitly oppose any such misuse. To mitigate this risk, we will release the dataset under an online usage agreement that requires all users to comply with applicable website terms and relevant legal and ethical standards, and that prohibits use intended to produce deceptive or harmful content.

All annotators of our dataset are coauthors of this manuscript rather than crowdworkers. They participated as invited volunteers (without using any third-party crowdsourcing platform), and the annotation process was conducted under our direct supervision.

References

Michele Bedard and Chianna Schoenthaler. 2018. Satire or fake news: Social media consumers' sociodemographics decide. In *Companion proceedings of the the web conference 2018*, pages 613–619.

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.

Sohan De Sarkar, Fan Yang, and Arjun Mukherjee. 2018. Attending sentences to detect satirical fake news. In *Proceedings of the 27th international conference on computational linguistics*, pages 3371–3380.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.

Hang Du, Guoshun Nan, Sicheng Zhang, Binzhu Xie, Junrui Xu, Hehe Fan, Qimei Cui, Xiaofeng Tao, and Xudong Jiang. 2024. Docmsu: A comprehensive benchmark for document-level multimodal sarcasm understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17933–17941.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. In *IJCAI*.

Rongpei Hong, Jian Lang, Jin Xu, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Following clues, approaching the truth: Explainable micro-video rumor detection via chain-of-thought reasoning. In *Proceedings of the ACM on Web Conference 2025*, pages 4684–4698.

Soumyadeep Jana, Animesh Dey, and Ranbir Singh Sanasam. 2024. Continuous attentive multimodal prompt tuning for few-shot multimodal sarcasm detection. In *Proceedings of the 28th Conference on Computational Natural Language Learning*, pages 314–326.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic sarcasm detection: A survey. *ACM Computing Surveys (CSUR)*, 50(5):1–22.

Manishit Kundu, Sumit Shekhar, and Pushpak Bhattacharyya. 2025. Looking beyond the pixels: Evaluating visual metaphor understanding in vlms. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23137–23158.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4707–4715.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, 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)*, volume 1, pages 1767–1777. Association for Computational Linguistics.

Ming Lu, Zhiqiang Dong, Ziming Guo, Xiaoming Zhang, Xinxi Lu, Tianbo Wang, and Litian Zhang. 2025. A multi-modal sarcasm detection model based on cue learning. *Scientific Reports*, 15(1):10261.

Abhijit Mishra, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya. 2016. Harnessing cognitive features for sarcasm detection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104.

Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. 2024. *** yesbut***: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16878–16895.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392.

Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 9507–9515.

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.

Victoria L Rubin, Niall Conroy, Yimin Chen, and Sarah Cornwell. 2016. Fake news or truth? using satirical cues to detect potentially misleading news. In *Proceedings of the second workshop on computational approaches to deception detection*, pages 7–17.

Binghao Tang, Boda Lin, Haolong Yan, and Si Li. 2024. Leveraging generative large language models with visual instruction and demonstration retrieval for multimodal sarcasm detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1732–1742.

Divyank Tiwari, Diptesh Kanojia, Anupama Ray, Apoorva Nunna, and Pushpak Bhattacharyya. 2023. Predict and use: Harnessing predicted gaze to improve multimodal sarcasm detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15933–15948.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.

Jie Wang, Yan Yang, Yongquan Jiang, Minbo Ma, Zhuyang Xie, and Tianrui Li. 2024a. Cross-modal incongruity aligning and collaborating for multi-modal sarcasm detection. *Information Fusion*, 103:102132.

Peng Wang, Yongheng Zhang, Hao Fei, Qiguang Chen, Yukai Wang, Jiasheng Si, Wenpeng Lu, Min Li, and Libo Qin. 2024b. S3 agent: unlocking the power of vllm for zero-shot multi-modal sarcasm detection. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Qiaofeng Wu, Wenlong Fang, Weiyu Zhong, Fenghuan Li, Yun Xue, and Bo Chen. 2025. Dual-level adaptive incongruity-enhanced model for multimodal sarcasm detection. *Neurocomputing*, 612:128689.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Ben Yao, Yazhou Zhang, Qiuchi Li, and Jing Qin. 2025. Is sarcasm detection a step-by-step reasoning process in large language models? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25651–25659.

Tan Yue, Xuzhao Shi, Rui Mao, Zonghai 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.

Qinglong Zhang and Yu-Bin Yang. 2021. Rest: An efficient transformer for visual recognition. volume 34, pages 15475–15485.

Yazhou Zhang, Chunwang Zou, Zheng Lian, Prayag Tiwari, and Jing Qin. 2025. Sarcasmbench: Towards evaluating large language models on sarcasm understanding. *IEEE Transactions on Affective Computing*.

Appendix

A Human Annotation Instructions

This appendix describes the human annotation protocol for satirical cartoons, including the operational definition of satire, the labeling rationale, and the practical workflow implemented with Label Studio.

A.1 Operational Definition of Satire

In this work, *satire* is defined as the use of rhetorical devices such as metaphor and exaggeration to expose, criticize, or ridicule a person, group, or social phenomenon. We adopt the definition from *Xinhua Dictionary* as the normative reference. For annotation, satire is treated as a semantic structure in which the surface meaning is intentionally misaligned with the intended meaning, commonly realized through irony, contrast, or visually grounded critique.

A.2 Labeling Scheme and Aggregation Rule

To reduce ambiguity during annotation, we employ three labels that explicitly distinguish non satirical content from satire with different degrees of directedness or hostility. This three class design is introduced primarily to make the annotator's decision process easier and more consistent. In the final dataset used for modeling and evaluation, we merge `label=1` and `label=2` into a single positive category, namely *satire*, while `label=0` remains the negative category.

When an instance is difficult to judge, annotators may consult an AI assistant by providing the definitions and diagnostic cues in this section as contextual information. The assistant's explanation is used only as a reference. The final label is always determined by the human annotator.

A.3 Label Definitions and Diagnostic Cues

Label 0: Non Satirical. An instance is labeled as non satirical when neither the image nor the title exhibits a satirical structure. In particular, there is no discrepancy between literal expression and intended meaning, no conflict between the title and the depicted content, no visual metaphor that implies a directed critique, and no irony. Typical instances show high consistency between the title and the image, and they do not present an identifiable target of criticism. Exaggeration may appear for artistic or humorous effect, and emotional tones such as resignation, light humor, or lyricism are allowed as long as they do not function as satire.

Label 1: Non Aggressive and Non Malicious Satire. An instance is labeled as non aggressive satire when a satirical structure is present but the instance does not attack a person or an entity. The intent is primarily self directed or situation oriented, such as self deprecating irony or a non confrontational portrayal of everyday predicaments. The instance typically lacks a specific target of attack, and its affect often reflects resignation, absurdity, or mild complaint rather than hostility.

Label 2: Aggressive and Malicious Satire. An instance is labeled as aggressive satire when a satirical structure is present and the instance conveys a directed intent to criticize, attack, or expose. This class is characterized by contradiction, tension, or metaphorical mismatch between the title and the image, and by a clearly identifiable target of critique. Common targets include governments, political systems, and bureaucratic structures; capital, corporations, and workplace exploitation; unequal allocation of social resources, class stratification, wealth inequality, and unequal opportunity; public services such as education, healthcare, elder care, and housing; technology, media, and digital life including smartphones, the internet, online platforms, and advertising; environmental and ecological issues such as pollution, performative environmentalism, ecological destruction, and resource

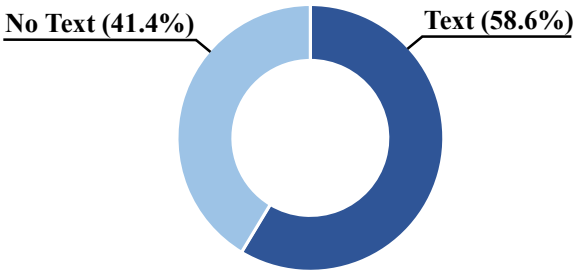

Figure 5: Percentage of images containing embedded text.

waste; war, violence, security, terrorism, public safety, and arms races; and collective social behavior and norms such as bandwagon effects, indifference, moral numbness, and conformity. Targets outside these categories are also allowed when the critique is explicit.

A.4 Annotation Workflow in Label Studio

The annotation environment is reproduced locally using Label Studio. Annotators install Label Studio in a VS Code terminal with Python version at least 3.8 by running `pip install label-studio`. They then start the service by running `label-studio` and keep the terminal session active during annotation. The interface opens at `http://localhost:8080`. Annotators create a local account on first use and log in, noting that all data are stored locally.

A new project is created with the name `sarcasm_detection`. The dataset file `clean_data.json` is uploaded through the *Data Import* entry. The labeling interface is configured by pasting the contents of `templates.txt` into the template editor and saving the configuration. Annotation is performed by entering the project and selecting *Label All Tasks*, assigning exactly one label to each instance according to the definitions and cues above. After all instances are labeled, the results are exported using the built in export function in the Label Studio interface.

B Dataset Analysis

Figure 5 summarizes the prevalence of embedded text in images: 58.6% of images contain on-image text, while 41.4% do not.

C Experiment Setup

C.1 Description of Baselines

In our experiments we compare against both multimodal sarcasm detection methods that are built on top of multimodal large language models (MLLMs) and those that are *not* based on MLLMs. The distinction is important: methods based on MLLMs exploit large pretrained vision–language reasoning capabilities to perform inference with minimal task-specific training, whereas the non-MLLM baselines typically rely on task-specific feature fusion and dedicated classifiers.

The baselines we evaluate include the following five representative approaches.

- **SarcasmCue (Yao et al., 2025)** investigates whether large language models can detect sarcasm via intermediate reasoning steps. It introduces a prompting framework with multiple strategies that elicit reasoning about linguistic and contextual cues and evaluates LLMs on standard sarcasm benchmarks. This study highlights that non-sequential prompting often outperforms sequential prompting even for LLMs, suggesting that sarcasm comprehension may not strictly follow a linear reasoning process.

- **S^3 Agent (Wang et al., 2024b)** proposes a multi-view agent framework that unlocks the power of vision large language models for zero-shot multimodal sarcasm detection. By decomposing sarcasm signals into superficial expression, semantic information, and sentiment views, this method leverages the reasoning ability of pretrained vision LLMs to achieve strong zero-shot performance on MMSD2.0.

- **DocMSU (Du et al., 2024)** introduces a comprehensive benchmark for document-level multimodal sarcasm understanding in news articles. In addition to the dataset, the authors propose a baseline approach that aligns pixel-level image features with word-level textual representations to capture sarcasm cues across long texts and heterogeneous content.

- **Dual-Level Adaptive Incongruity-Enhanced model (DAIE) (Wu et al., 2025)** is a deep architecture emphasizing multimodal incongruity. It encodes text and visual modalities and then fuses information at token and graph levels to explicitly highlight contradiction patterns, which are central to multimodal sarcasm detection.

- **MMSD2.0 (Qin et al., 2023)** defines a corrected multimodal sarcasm detection system and benchmark that addresses issues in the original MMSD dataset by removing spurious cues and reannotating unreasonable samples. A multi-view CLIP framework is introduced to fuse text, image, and text–image interaction representations, substantially improving reliability and baseline performance on this benchmark.

- **Twitter Hierarchical Fusion Model (MMSD)** (Cai et al., 2019) represents an early multimodal sarcasm detection approach on social media, using hierarchical fusion of text and image features with attention mechanisms. This model integrates modality features at multiple levels to highlight cross-modal inconsistencies for sarcasm detection, though without relying on MLLMs.

Among these baselines, the approaches that are based on multimodal large language models are S^3 Agent and SarcasmCue, because both take advantage of large pretrained vision or language models to make predictions with little or no task-specific training. SarcasmCue investigates how internal reasoning processes in large language models relate to sarcasm understanding, and S^3 Agent uses vision LLMs to address the multimodal sarcasm detection problem in a zero-shot scenario. The other baselines such as the alignment baseline in DocMSU, the dual-level adaptive incongruity model, the MMSD2.0 multi-view CLIP framework, and the hierarchical fusion model for Twitter focus on customized multimodal representations and feature fusion rather than on large model reasoning. Taken together, these methods provide a comprehensive set of comparative systems for evaluating multimodal sarcasm detection performance.

D Implementation Details

In this section, we present detailed implementation specifications for our proposed DOUBLE as well as a comprehensive overview of the experimental setup.

D.1 Implementation Environment

All experiments are conducted on a system comprising an Intel(R) Core(TM) i9-14900KF processor, equipped with one NVIDIA GeForce RTX 4090 D GPU with 24 GB of VRAM, and accompanied by 64 GB of DRAM.

Table 4: Example of prompt for various perspectives applied in S^3 Agent.

Superficial Expression Prompt
Given the following image and text, judge whether there is sarcasm based on the superficial expression. This requires detecting underlying critique through discrepancies between image and text, considering both modalities together.
`text: <text>`

Semantic Information Prompt
Given the following image and text, judge whether there is sarcasm based on the semantic information. This prompt focuses on extreme portrayals and metaphorical meanings inferred from the combined image and text.
`text: <text>`

Sentiment Expression Prompt
Given the following image and text, judge whether there is sarcasm based on sentiment expression. This requires detecting critical emotions targeted at specific subjects or behaviors by jointly examining the image and text.
`text: <text>`

D.2 MLLM Implementation in Clue2View

To ensure the robustness and generalizability of our findings, we conduct all experimental evaluations using a diverse set of three state-of-the-art MLLMs: GPT-4o-mini, GPT-5.1, and Qwen2.5-VL-72B-Instruct. By employing this range of models, we verify the effectiveness of our approach across different architectural designs and resource constraints."

D.3 SLM Implementation in SLM Arbiter

Our SLM is based on a masked self-attention Transformer architecture, i.e., BERT, pretrained through language-visual contrastive learning (Radford et al., 2021). For the visual feature encoding, we leverage the pre-trained Vision Transformer (ViT), keeping its parameters frozen. Specifically, for our benchmarks, we employ OFA-Sys/chinese-clip-vit-large-patch14 (Yang et al., 2022).

D.4 Data Preprocessing

Our SLM is based on a masked self-attention Transformer architecture, i.e., BERT, pretrained through language-visual contrastive learning (Rad-

ford et al., 2021). For the visual feature encoding, we leverage the pre-trained Vision Transformer (ViT), keeping its parameters frozen. Specifically, for our benchmarks, we employ OFA-Sys/chinese-clip-vit-large-patch14 (Yang et al., 2022).

Given an image–text pair X, we first extract its multimodal information from both modalities. For the visual content, we represent the image as I and adopt the vision encoder of `OFA-Sys/chinese-clip-vit-large-patch14` (Yang et al., 2022) as the feature extractor. Specifically, we feed I into the CLIP vision transformer and take the output embedding of the `[CLS]` token as the global image representation, denoted as $\mathbf{h}_v \in \mathbb{R}^{d_v}$, where d_v is the dimensionality of the visual embedding space defined by the model.

For the textual content, we use the image associated title and external background knowledge as the textual input. Specifically, we denote the title as $T_{\text{title}} \in \mathbb{R}^{n_t}$ and the background knowledge as $T_{\text{bg}} \in \mathbb{R}^{n_b}$, where n_t and n_b are the numbers of words (or tokens) in the title and the background text, respectively. We then concatenate them to form the final textual sequence $\tilde{T} = [T_{\text{title}}; T_{\text{bg}}] \in \mathbb{R}^{\tilde{n}}$, where $\tilde{n} = n_t + n_b$. This design enables the model to interpret the image in context by jointly considering the visual content, the title, and the relevant background information.

D.5 Training Configuration

Experimental Setup. We implement our model using the PyTorch framework. For text encoding, we utilize the pre-trained Chinese-CLIP-ViT-Large-Patch14 model to extract textual features. For visual features, we align the input dimensions through a multi-layer perceptron (MLP) projection. We employ the AdamW optimizer for model parameter optimization, configured with a learning rate of 7.7×10^{-5} and a weight decay of 5×10^{-4}. To mitigate the class imbalance issue, we adopt the Focal Loss with γ set to 1.8. The maximum number of training epochs is set to 15, with an early stopping mechanism triggered if the validation accuracy does not improve for 5 consecutive epochs. The batch size is set to 32, and the random seed is fixed at 2026 to ensure reproducibility. For statistical reliability, we report the results based on the best-performing model selected on the validation set. For baseline models, we strictly adhere to the settings specified in their original papers.

Prompt$_1$: Literal Scene Graph

Role: You are designated as a specialized agent for Multimodal Knowledge Graph construction, focusing on surface-level visual and textual grounding. The extraction process is divided into two sequential phases: (A) Initial Multimodal Parsing and (B) Surface-KG Refinement.
Phase A: Fine-Grained Multimodal Parsing Objective: Perform a comprehensive extraction of entities, attributes, and relationships derived solely from the visual content and the provided title metadata. Methodology: Employ multi-scale visual inspection (zooming or cropping) to resolve fine-grained details, particularly for textual elements, symbols, or small objects. Inference Constraint: All extractions must be strictly grounded in observable evidence. Probabilistic inference is permissible only when strongly supported by visual cues; hallucination of unobserved entities, relations, or events is strictly prohibited.
Phase B: Surface-KG Refinement Objective: Critique the initial extraction to produce a finalized, revised Surface Knowledge Graph. This graph must contain only factual assertions directly verifiable in the image, strictly filtering out over-interpreted semantic abstractions.
Operational Protocols and Constraints:

1.Entity Extraction Protocols: Taxonomy: All entity nodes must be labeled in English. Granularity: Decompose composite objects into sub-entities when distinct interactions (e.g., handling, using) are observed. Textual Entities: Extract all visible text strings as distinct entities. Preserve non-English content exactly as it appears within attribute values, but maintain English node labels.

2.Relation Extraction Protocols: Predicate Ontology: Relations must be expressed as specific, English action verbs or spatial prepositions. Grounded Triples: Generate only factual triplets (Entity-Relation-Entity) substantiated by observable spatial arrangements, actions, interactions, or communicative gestures. Speaker-Utterance Alignment: For dialogue or text associated with a specific character, enforce the structure: [Person_Entity] -> [say] -> [Text_Entity].

3.Attribute Extraction Protocols: Visual Attributes:Annotate entities with visually grounded descriptors including color, material, pose, facial expression, relative size, and spatial coordinates. Textual Attributes: For on-screen text, include attributes such as raw_text_content, layout_structure, reading_direction, and text_type.

4.OCR and Layout Analysis (Chinese/Multilingual): Segmentation: Extract visible text fragments as individual entities. Reconstruction: Merge text fragments only when the layout strongly implies continuity; otherwise, treat them as discrete. If the reading order is ambiguous, annotate the direction as 'unknown'.

5.Evidence Modality Annotation: Requirement: Every relation and attribute in the final output must include an 'evidence_modality' tag to specify the information source. Valid Tags: 'image' (visual only), 'text' (textual only), or 'image+text' (cross-modal inference).

Input: Image and Title. Output: A structured, purified Surface Knowledge Graph adhering to the above constraints.

(a) Step 1

Prompt$_2$: Non-Sarcastic Interpretation

You are an expert in abductive reasoning. Your task is to analyze the provided {surface_kg} to generate a streamlined rationale from a literal perspective.
Strictly adhere to the provided Knowledge Graph nodes and edges without hallucinating external information.
Task:
Using ONLY the {surface_kg}, explain why the sample appears reasonable or non-sarcastic in its literal sense. Focus on the visible facts and standard interpretations of the objects and actions described.
Please answer in Chinese.
Return ONLY valid JSON with the following structure:

```
{
 "non_sarcastic_explanation": "string"
}
```

(b) Step 2

Prompt$_3$: Refined Scene Graph

Role: You are designated as a specialized agent for the construction of Deep Semantic Knowledge Graphs (Deep-KG), specifically tailored for the interpretation of satirical cartoons. Your objective is to model the latent semantic layer—symbolism, intent, and causal logic—essential for computational sarcasm detection.
Input Specifications:

Meta-Information: Title text and optional background context.
Visual Narrative: A concise LLM-generated description of the scene.
Surface Knowledge Graph: A structured representation of explicitly visible facts.

Methodological Framework:
Semantic Abstraction and Inference: Construct a Deep-KG that captures implied meanings, metaphorical mappings, social roles, causal dependencies, communicative intents, normative expectations, and abstract concepts. The graph must bridge the gap between visual signals and high-level interpretation.
Redundancy Constraint: Strictly exclude purely visible facts already present in the Surface KG, unless they serve as necessary semantic anchors for connecting abstract nodes (e.g., grounding a symbol to its physical representation).
Node Ontology and Naming: You are authorized to introduce abstract nodes where supported by the title, background context, or scene composition. All node identifiers must adhere to snake_case formatting. Existing node IDs should be preserved where applicable; new nodes should be generated only to represent distinct semantic concepts.
Output Specifications: Return the result strictly as a valid JSON object adhering to the schema defined below. Do not include markdown formatting, code block delimiters, or explanatory text.

(c) Step 3

Prompt$_4$: Sarcastic Interpretation

You are an expert in abductive reasoning for sarcasm detection. Your task is to analyze the provided {surface_kg} and {deep_kg} to generate a streamlined rationale from a sarcastic perspective. Strictly adhere to the provided Knowledge Graph nodes and edges without hallucinating external information.
Task:
Combining the {surface_kg} and {deep_kg}, explain why the sample is reasoned as sarcastic. Identify semantic conflicts, common sense violations, or contradictions between the surface layer (what is seen) and the deep layer (symbolism, intent, or cultural context).
Please answer in Chinese.
Return ONLY valid JSON with the following structure: { "sarcastic_explanation": "string" }

(d) Step 4

Figure 6: Prompting overview.