MMSD3.0: A Multi-Image Benchmark for Real-World Multimodal Sarcasm Detection

Supplementary Material

Table 5. MMSD3.0 results under the modified protocol: per-image encoding with feature concatenation (no canvas tiling). Best results are in **bold**.

Modality	Method	MMSD3.0			
		Acc (%)	**P** (%)	**R** (%)	**F1** (%)
Image-Only	ResNet [14]	65.33	53.13	36.18	43.05
	ViT [9]	64.03	50.31	55.12	52.61
Multimodal	DIP [42]	81.49	81.17	79.95	80.41
	Multi-view CLIP [34]	82.20	75.86	79.81	77.79
	MoBA [43]	80.17	79.26	78.83	79.02
	Tang et al. [38]	82.39	81.47	81.61	81.54
Ours	**CIRM**	**85.16**	**84.41**	**84.43**	**84.42**

Table 6. Comparison of model performance on real-world and AI-generated data.

Modality	Method	Real-world		AI-Gen
		Acc (%)	**F1** (%)	**Acc** (%)
Multimodal	DIP [42]	79.59	75.50	97.98
	Multi-view CLIP [34]	80.01	75.48	95.96
	MoBA [43]	76.02	68.80	88.89
	Tang et al. [38]	80.36	75.93	95.45
MLLM	GPT-4o [30]	71.34	67.60	82.61
	LLaVA-1.5-7B [24]	59.62	55.78	72.83
	Qwen2.5-VL-32B [1]	68.90	66.65	95.65
Ours	**CIRM**	**83.31**	**80.39**	**98.48**

5.8. No-Tiling Multi-Image Experiment

Setting. Because concatenating images can introduce unfairness, we further examine how mainstream methods perform when the framework is adapted to accept multiple images as input. Specifically, we no longer concatenate images into a single composite image. Instead, we encode each image and then concatenate the resulting feature vectors. We modify only the input format, and the model architecture remains unchanged.

Baselines. *CIRM* and modern MLLMs natively accept multiple images; single-image baselines are adapted by concatenating per-image features (no canvas tiling). **Image-only:** ResNet [14] (residual CNN) and ViT [9] (patch-token Transformer), both using per-image encoding with feature concatenation (no canvas tiling). **Multimodal (representative):** DIP [42] (joint factual/affective fusion); Multi-view CLIP [34] (multi-perspective CLIP alignment); MoBA [43] (bidirectional modality interaction for robust fusion); Tang et al. [38] (LLM prompting with retrieved demonstrations).

Fairness & Reproducibility. We use official hyperparameters when available; trained baselines follow the same data splits and evaluation protocol as *CIRM*.

Main Results. Under the modified protocol that encodes each image separately and concatenates the resulting features, **CIRM** continues to achieve the highest performance on MMSD3.0 with 85.16 Acc and 84.42 F1. Image-only baselines such as ResNet and ViT remain weak, showing that visual cues alone are insufficient for sarcasm understanding. Among multimodal methods, Tang et al. and DIP retain relatively strong results, while Multi-view CLIP and MoBA show smaller gains or even reduced effectiveness, indicating that simply aggregating independent features cannot model multi-image dependencies effectively. Overall, these findings confirm that late feature concatenation provides limited benefit compared with architectures that perform explicit cross-image reasoning. **CIRM** sustains its advantage because its dual-bridging mechanism captures both inter-image relationships and cross-modal alignment, leading to a more complete multimodal understanding.

5.9. Real-World Performance

As illustrated in Table 6, there is a significant drop in model performance on real-world data compared to AI-generated samples, which underscores the inherent challenges posed by the greater complexity and unpredictability of real-world posts. The variation in context, language, and multimodal cues contributes to this discrepancy. Among the multimodal models, DIP and Tang et al. stand out with relatively competitive performance, whereas MoBA underperforms, and Multi-view CLIP exhibits limited adaptability to real-world conditions. On the MLLM front, while GPT-4o and Qwen2.5-VL-32B deliver moderate accuracy, LLaVA-1.5-7B lags behind in both accuracy and F1 score.

In contrast, **CIRM** consistently outperforms all baselines, achieving a robust F1 score of 80.39% on real-world data and a remarkable 98.48% on AI-generated samples. This reinforces the model's versatility and highlights the effectiveness of its cross-image reasoning and Relevance-Guided Fusion strategies. By focusing on cross-modal consistency and latent inter-image relationships, CIRM demonstrates superior capacity to capture and interpret the subtle, multifaceted nature of sarcasm in real-world multimodal

Table 7. Paired truncation on MMSD3.0 long-text ($L \geq 30$, n=500) with CIRM. Δ is Full − Trunc (pp).

View	Acc [95% CI]	P	R/F1
Full	86.40 [83.12, 89.13]	84.67	83.71 / 84.16
Trunc	73.40 [69.36, 77.08]	69.43	65.22 / 66.18
Δ	**+13.00**	**+15.24**	**+18.49 / +17.98**

data. The results not only confirm the importance of tailored cross-image reasoning but also establish CIRM as a promising solution for multimodal sarcasm detection in complex, varied environments.

5.10. Length Sensitivity via Paired Truncation

Setup. We reuse the same *CIRM* checkpoint as in the main experiments and evaluate on a randomly sampled long text subset of MMSD3.0 with $L \geq 30$ tokens ($n = 500$). Each item is fed to the model under two input views. *Full* keeps the original text. *Trunc* keeps the first 15 tokens and preserves emojis and punctuation. Model parameters remain fixed so that the only difference arises from input length.

Metrics. We report Accuracy and Macro Precision, Recall, and F1. Accuracy includes the Wilson 95% confidence interval to indicate estimation uncertainty.

Results. Table 7 shows a clear separation between the two views. *Full* attains 86.40% Accuracy with a 95% confidence interval of [83.12, 89.13], and 84.67% Macro Precision, 83.71% Macro Recall, and 84.16% Macro F1. *Trunc* attains 73.40% Accuracy with a 95% confidence interval of [69.36, 77.08], and 69.43% Macro Precision, 65.22% Macro Recall, and 66.18% Macro F1. The differences computed as Full minus Trunc are +13.00 percentage points on Accuracy, +15.24 on Macro Precision, +18.49 on Macro Recall, and +17.98 on Macro F1. The confidence intervals for Accuracy do not overlap, which indicates a statistically meaningful gap.

Interpretation. Truncation removes tail spans that often contain decisive sarcasm cues such as contrastive or negated clauses, quoted punchlines, and trailing emojis or OCR referenced phrases. Because items and weights are identical across views, the observed degradation is attributable to the loss of long range textual evidence rather than training or sampling differences. The larger drops in Recall and F1 suggest that many positive instances become harder to recognize once late cues are removed, while the decline in Precision indicates increased ambiguity in shortened inputs. This test supports the claim that longer texts in MMSD3.0 supply essential information for robust multimodal sarcasm detection and that the proposed model effectively exploits extended context.

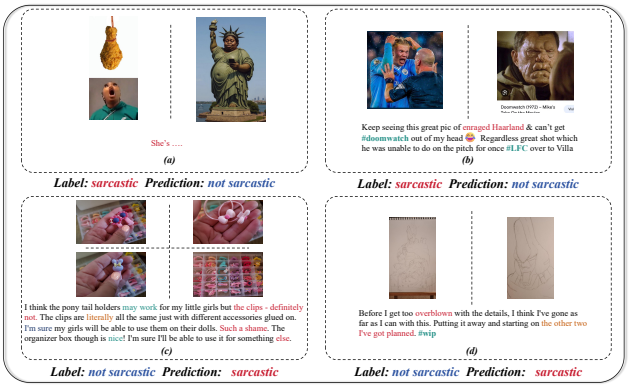

Figure 7. Error cases illustrating challenging sarcastic scenarios.

5.11. Error Analysis

Figure 7 illustrates representative failure cases revealing key challenges in multi-image sarcasm detection. In (a) and (b), the model overlooks abstract or culture-dependent incongruity. Example (a) fails mainly because the text is too short, providing no contextual signal for interpreting the visual metaphor. The sarcastic intent relies on exaggerated visual analogy across images, which the model misses without textual grounding. In (b), the model fails to recognize sarcasm connected to meme-style humor and football culture, where the reference to "doomwatch" depends on shared cultural knowledge that the model cannot infer from surface text.

In (c) and (d), subtle linguistic pragmatics—phrases like "may work," "such a shame," or "too overblown"—carry mixed emotional tones that confuse the sentiment classifier. These ambiguous expressions blend positive and negative cues, making it difficult for the model to correctly detect sarcasm. Overall, these cases indicate that beyond multimodal fusion, sarcasm detection requires commonsense reasoning, contextual grounding, and deeper sensitivity to pragmatic and cultural nuances.

Prompt Template for Generating Tweets

```
You are a real-world Twitter user.
You will be given multiple related images.
Write 3 different English sarcastic tweets.
Requirements:
- At least one tweet must use the relationships, contrasts,
  or narratives across the images (multimodal sarcasm).
- Others may focus on a single image or general sarcastic styles
(verbal irony, exaggeration, understatement, sarcastic comparison).
- Tweets should be natural.
- Hashtags (#) and/or emojis may appear naturally in any position
  of the tweet (beginning, middle, or end not only at the end), and
  multiple hashtags/emojis may be used if suitable.
- Output exactly 3 tweets in this format, with no explanations:
1: <tweet>
2: <tweet>
3: <tweet>
```

Figure 8. Prompt Template for Generating Tweets.

Prompt Template for Generating Reviews

```
You are a real-world Amazon customer writing product reviews.
You will be given multiple related product images.
Write 3 different English sarcastic reviews for this product.
Requirements:
- At least one review must use the relationships, contrasts,
or narratives across the images (multimodal sarcasm).
- Others may focus on a single image or general sarcastic styles
(verbal irony, exaggeration, understatement, sarcastic comparison).
- Reviews should be natural, fluent, and realistic
 as if posted on Amazon.
- Emojis may appear naturally anywhere (beginning, middle, or end)
 if suitable, and multiple emojis may be used.
- Output exactly 3 reviews in this format, without explanations:
1: <review>
2: <review>
3: <review>
```

Figure 9. Prompt Template for Generating Reviews.

Prompt Template for Evaluating

```
You are an expert in evaluating user-generated content for {platform}.
You will be provided with three sarcastic texts generated based on a set of related images.
Your task is to evaluate these texts and select the best one based on the following criteria:

1. Naturalness and Authenticity: The text should feel like it was written by a real user on the {platform}.
It should match the tone, style, and conventions of typical posts or reviews on that platform
(e.g., concise and casual for Twitter, detailed and product-focused for Amazon).

2. Effectiveness of Sarcasm: The sarcasm should be sharp, clear, and impactful,
using techniques like verbal irony, exaggeration, understatement, or sarcastic comparison.
It should be humorous or critical without being overly mean-spirited or forced.

3. Image Integration: For texts that reference multiple images, the sarcasm should effectively
leverage relationships, contrasts, or narratives across the images (multimodal sarcasm).
For texts focusing on a single image, the sarcasm should align well with the visual content.

4. Coherence and Relevance: The text should be coherent, contextually relevant to the images,
and avoid feeling abrupt, out-of-place, or unnatural.

5. Platform-Specific Fit: For tweets, ensure the text is concise
(within Twitter's character limit of 280 characters), engaging, and uses hashtags/emojis naturally if present.
For Amazon reviews, ensure the text is appropriately detailed, product-focused, and mimics the style of real
customer reviews.

6. Subtlety and Creativity: The best text should strike a balance between being obviously sarcastic and
 subtle enough to feel clever and realistic, avoiding clichés or overly exaggerated humor.

Platform: {platform}

Generated sarcastic texts:{generated_text}

Please respond with the BEST sarcastic text from the above generated texts,
followed by your reasoning (max 100 words) for why it's the best choice.

Format:
BEST_TEXT: [paste the exact best text here without any modifications]
REASON: [your explanation]
```

Figure 10. Prompt Template for Evaluating

Prompt Template for MLLMs Sarcasm Detection

```
Please analyze the following multimodal content (text and images)
objectively to determine whether it contains sarcasm.
Evaluate the intent behind the text and the visual elements together,
considering contextual cues, tone, and contrast between literal meaning
and implied sentiment.

The images are presented in their original order,
which may be important for understanding the context.

Respond strictly with:

"1" if the content is sarcastic

"0" if the content is not sarcastic

Provide no additional explanation.

Text: {text}

Answer:
```

Figure 11. Prompt Template for MLLMs Sarcasm Detection