CAFE: Caption-Anchored Feature Enrichment and Evidence Routing for Multimodal Sarcasm Detection*

Puchu Zeng[1,†] **Yun Luo**[2,†] **Qianqian Li**[3]
Danhui Yan[2,‡] **Liu Junyi**[2] **Fengxuan Song**[2]

[1]Hunan Normal University, Changsha, Hunan, China
[2]National University of Defense Technology, Nanjing, Jiangsu, China
[3]University of Science and Technology of China, Hefei, China

[†]Equal contribution. [‡]Corresponding author.

Abstract

Multimodal sarcasm detection aims to identify sarcasm in image–text posts by capturing how visual content changes the interpretation of short informal text. Existing methods have advanced the task through multimodal fusion, caption-based cues, and fact/sentiment incongruity modeling, but their evidence use is often controlled inside latent feature channels or reduced discrepancy scores. This makes it difficult to decide which visual information should be selectively weighted for a particular post. We propose CAFE, a caption-anchored feature enrichment and evidence routing framework that decides how to use visual evidence for each sample under ambiguity. CAFE enriches each image with two generated descriptions—one for visible facts and one for the image–text relation—then estimates how useful each description is for the current post and fuses caption-calibrated contrast evidence with direct cross-modal evidence for prediction. Experiments show that CAFE achieves state-of-the-art performance on the primary MMSD2.0 benchmark with 87.81±0.33% F1 (mean over 5 seeds). Component ablations on the original MMSD provide an additional evaluation.

1 Introduction

Image–text posts on social media often convey meaning through the interaction between visual content and short informal text. In such posts, the image may provide context, irony, or an implicit stance that is not explicitly stated in the text. Multimodal sarcasm detection therefore requires more than recognizing visual objects or textual sentiment; it requires judging how visual evidence changes the interpretation of the post.

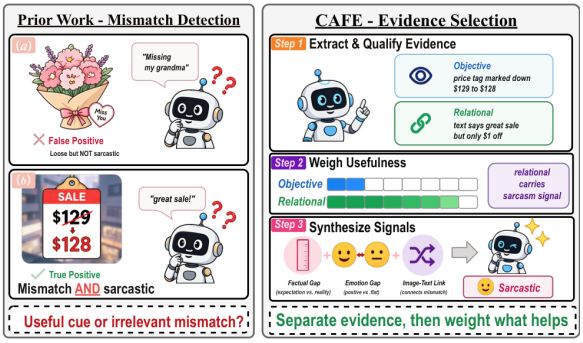


Figure 1: Evidence flow in CAFE: grounded generation produces factual and relational descriptions, utility calibration estimates their sample-specific usefulness, and the detector combines calibrated evidence with cross-modal cues.

A central challenge is that image–text mismatch is not equivalent to sarcasm. Many non-sarcastic posts contain decorative, humorous, or loosely related images, where the visual content may appear weakly aligned with the text but does not express a sarcastic intent. Conversely, sarcastic posts often rely on a specific visual cue that reverses, reframes, or sharpens the literal meaning of the text. Thus, the difficulty is not merely to measure whether the image and text are discrepant, but to determine whether a visible cue provides decision-relevant pragmatic evidence for the current post. A decorative image, a humorous meme template, and a sarcasm-bearing visual cue may all appear weakly aligned with the text, but they require different treatment by the detector.

Existing methods can be broadly divided into two progressive categories. Early works adopt multimodal fusion methods, which combine textual

*Warning: This paper contains real social-media examples of sarcasm that may include offensive language or sensitive topics. They are presented solely for scientific illustration and do not reflect the views of the authors.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3103–3119
November 6-10, 2026 ©2026 Association for Computational Linguistics

features, visual features, and image attributes to exploit visual context beyond text-only sarcasm detection (Schifanella et al., 2016; Castro et al., 2019; Cai et al., 2019). These methods demonstrate that images are useful for sarcasm recognition; however, the fused representation is usually implicit and does not indicate which visual contents are actually useful for a given prediction. To address this, recent works propose incongruity-aware methods that add view-specific channels—factual, affective, topological, or graph-structured—on top of the encoder and combine them through routing, graph reasoning, tension fields, or multi-view contrastive learning (Tian et al., 2023; Zhu et al., 2024; Gao et al., 2025; Zhang et al., 2025; Guo et al., 2025; Li et al., 2025). These methods have substantially improved cross-modal representation learning by measuring richer forms of mismatch. At the same time, many of them operate after heterogeneous visual evidence has already been compressed into latent channels, interaction vectors, or discrepancy features. We argue that understanding multimodal sarcasm also requires an earlier evidence qualification step: before the detector relies on a mismatch signal, it should assess what role the underlying visual evidence plays in the current post and how reliable that evidence is for the decision.

This view suggests that multimodal sarcasm detection is not only a representation problem but also an *evidence-routing* problem: given a concrete post, the model has to decide *which* piece of visual evidence is worth attending to for *this* decision, before the evidence is combined. A natural way to make visual evidence easier to compare with the post text is to turn the image into generated descriptions, and recent caption-based methods have already shown that such descriptions can stabilize cross-modal comparison (Tang et al., 2024; Aggarwal et al., 2024; Chen et al., 2024; Zhang et al., 2026). In these systems, generated text is typically used to support discrepancy estimation or representation enrichment after a description has already been selected as the visual proxy. This leaves less room to ask an earlier question: what role should a generated description play for this particular post, and how reliable is it as evidence before it is compressed into downstream comparison features? The key missing step is therefore earlier than final feature fusion: before combining evidence, the model should keep factual visual evidence and post-conditioned relational evidence available as separate textual candidates, then decide how much each one should influence the prediction for the current post.

To this end, we propose **CAFE**, a **C**aption-**A**nchored **F**eature **E**nrichment and evidence routing framework for multimodal sarcasm detection. CAFE first constructs two linguistically comparable evidence candidates: a grounded factual description of visible content and a post-conditioned relational description of how the image may affect the interpretation of the text. A role-aware utility module then evaluates these candidates in the context of the original post and forms calibrated description evidence, which is finally combined with lightweight contrast diagnostics and direct image–text interaction for prediction. The two design choices work together: role-separated descriptions make alternative forms of visual evidence explicit, and sample-wise calibration lets the model adjust their influence for different posts. Experiments on MMSD2.0 as the primary benchmark, with additional component ablations on the original MMSD, show that CAFE achieves state-of-the-art performance.

Our contributions are as follows: (1) We propose CAFE, a caption-anchored feature enrichment and evidence routing framework that treats generated descriptions as role-specific evidence candidates before fusion. (2) We introduce a role-aware utility calibrator that weighs each candidate by its sample-specific usefulness, enabling adaptive evidence qualification under ambiguity. (3) CAFE achieves state-of-the-art performance on MMSD2.0, with additional component ablations on MMSD, and the analyses show that role separation and sample-wise routing both contribute to the final detector.

2 Related Work

Multimodal sarcasm detection. Recent advances in multimodal sarcasm detection have shown that sarcastic intent in social media posts often depends on the joint interpretation of short text and accompanying images. Early studies established the task by collecting image–text sarcasm data and combining textual, visual, and image-attribute features in multimodal fusion models (Schifanella et al., 2016; Castro et al., 2019; Cai et al., 2019). The field then shifted toward stronger cross-modal reasoning: later methods model semantic association and cross-modal contrast (Xu et al., 2020), build graph-based or knowledge-enhanced interaction structures (Liang et al., 2022;

Liu et al., 2022), and introduce dynamic routing or debiasing mechanisms for noisy social-media posts (Tian et al., 2023; Jia et al., 2024; Wei et al., 2024). More recent systems further strengthen incongruity modeling with dual-granularity learning, topology-aware perception, ambiguity-aware fusion, and tension-field or multi-view incongruity representations (Zhu et al., 2024; Gao et al., 2025; Li et al., 2025; Zhang et al., 2025; Guo et al., 2025). Dataset and evaluation protocols have also become more reliable: MMSD2.0 corrects noisy labels and standardizes a stronger evaluation setting for image–text sarcasm detection (Qin et al., 2023). These methods make visual information increasingly comparable with text by learning richer interaction and incongruity representations. CAFE follows the same goal, but treats the reliability and role of textualized visual evidence as a sample-conditioned decision that should be made before such evidence is collapsed into final detector features.

Caption generation for visual evidence enrichment. The development of vision-language models has made generated descriptions a practical way to enrich visual evidence. Early image captioning models such as "Show and Tell" and "Show, Attend and Tell" used encoder–decoder architectures to convert visual features into natural language, while bottom-up/top-down attention improved the grounding of generated descriptions in salient image regions (Vinyals et al., 2015; Xu et al., 2016; Anderson et al., 2018). Vision-language pretraining then provided stronger multimodal encoders and contrastive alignment models, including cross-modal transformers, CLIP-style models, and generative vision-language models (Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2019; Chen et al., 2020; Radford et al., 2021; Li et al., 2022, 2023; Alayrac et al., 2022; Liu et al., 2023, 2024a). This capability has recently been introduced into multimodal sarcasm detection: generated visual descriptions, visual instructions, or LMM rationales are used as additional context, explanation signals, or anchors for discrepancy measurement (Tang et al., 2024; Aggarwal et al., 2024; Chen et al., 2024; Zhang et al., 2026). For example, factual captions can provide a stable semantic proxy for computing semantic or sentiment divergence between the image and the post text, as well as the fidelity of the generated description to the visual content. This line of work validates the usefulness of textualized visual evidence. CAFE follows this broader motivation, but keeps generated evidence in two role-specific textual forms—visible facts and post-conditioned relations—until a sample-conditioned utility module estimates how much each form should contribute.

Visual grounding and evidence routing. Although large vision-language models can describe images, generated descriptions may still hallucinate or over-interpret subtle social-media images. Visual grounding provides an effective way to constrain such generation by exposing object-level evidence before natural-language description. Image tagging models such as RAM and open-set detectors such as GroundingDINO can provide salient visual entities and localized object evidence for downstream vision-language processing (Zhang et al., 2024; Liu et al., 2024b). In parallel, sarcasm-specific incongruity models have shown that not every image–text mismatch is useful: the mismatch must be relevant to the implied attitude of the post. Recent work therefore introduces finer factual, affective, relation-aware, or memory-enhanced signals beyond global similarity (Tian et al., 2023; Lu et al., 2024; Zhang et al., 2025; Guo et al., 2025; Chen et al., 2026). However, these methods typically aggregate evidence through latent channels, attention weights, scalar discrepancies, or fixed fusion structures. CAFE combines grounded description enrichment with evidence routing: grounding constrains what is described, role-separated generation makes alternative evidence forms explicit, and utility calibration controls how those forms enter the final detector for each sample.

3 Method

3.1 Framework Overview

To decide which visual information should be used for each multimodal sarcasm post, we propose **CAFE**, a caption-anchored feature enrichment and evidence routing framework for multimodal sarcasm detection. CAFE generates two image descriptions, weighs them for the current post, and combines calibrated description evidence with direct image–text interaction and lightweight diagnostic cues. The overall architecture is shown in Figure 2.

Formally, the training set is denoted as

$$\mathcal{D} = \{(I_i, T_i, y_i)\}_{i=1}^{N}, \tag{1}$$

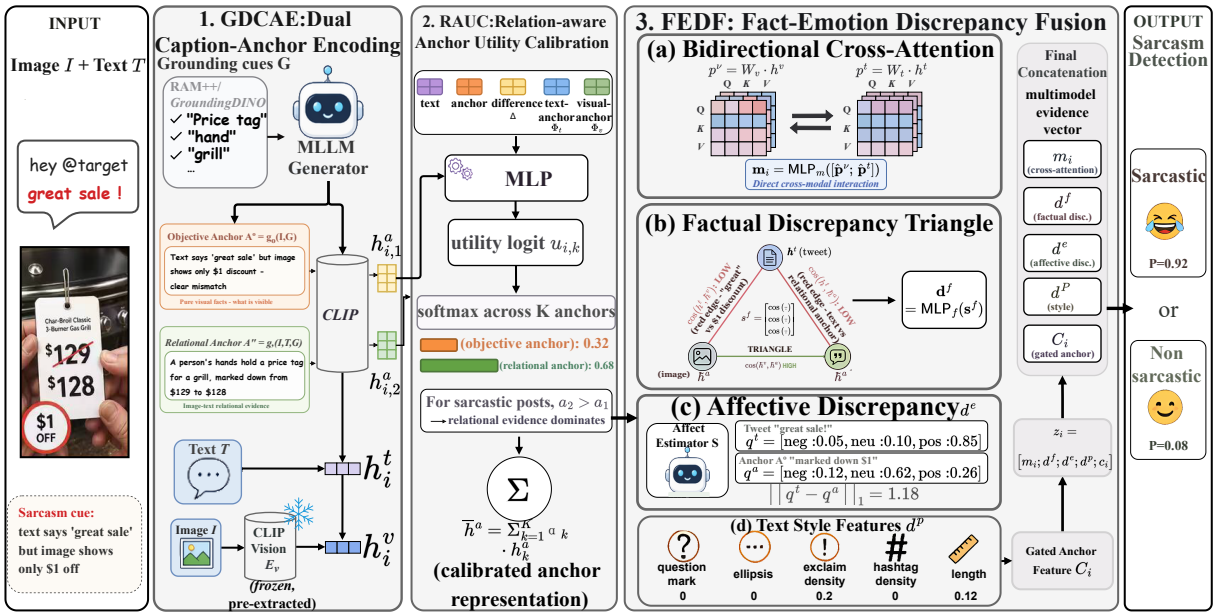


Figure 2: Overview of the CAFE framework. Given an image–text post, CAFE generates two descriptions—one for visible facts and one for the image–text relation—estimates their per-sample utility via the RAUC module, and combines the calibrated description evidence with direct cross-modal interaction and compact diagnostic cues for sarcasm prediction.

where I_i is the image, T_i is the post text, and $y_i \in \{0, 1\}$ is the sarcasm label. For each sample, CAFE constructs two generated descriptions, which we denote as anchors in the model:

$$\mathcal{A}_i = \{A_i^o, A_i^r\}, \tag{2}$$

where A_i^o is a factual image description and A_i^r is an image–text relation description. The model learns a function

$$f_\theta(I_i, T_i, \mathcal{A}_i) \rightarrow \hat{y}_i, \tag{3}$$

where $\hat{y}_i$ is the predicted sarcasm probability. Throughout this section, h_i^v and h_i^t denote the image and text representations, and $h_{i,k}^a$ denotes the representation of the k-th generated description. In the utility calibration module, $u_{i,k}$ is the utility score for that description, $\alpha_{i,k}$ is its normalized weight, and $\bar{h}_i^a$ is the weighted description representation. In the fusion module, m_i denotes the cross-modal interaction feature, d_i^f and d_i^e denote compact factual and affective diagnostic features, and z_i denotes the final feature used for sarcasm prediction.

Figure 2 illustrates the overall framework. CAFE follows a grounded enrichment and routing pipeline for sarcasm detection. First, **Grounded Dual Caption-Anchor Evidence Encoding (GDCAE)** uses detected visual entities as grounding context to generate two label-free textual evidence candidates: one for visible facts and one for the image–text relation. The grounding context and prompts are used only during offline generation, while the detector receives only the final generated sentences. Second, **Relation-aware Anchor Utility Calibration (RAUC)** acts as a soft evidence filter: it checks how useful each description is for the current image–text pair and forms a weighted description representation. Third, **Caption-Calibrated Contrast Evidence Fusion (CCEF)** combines the calibrated description evidence, the original image–text interaction, and compact post-calibration diagnostics to predict whether the post is sarcastic. The **Label-Smoothed Sarcasm Prediction Objective (LSSPO)** is described in Section 3.5.

3.2 Grounded Dual Caption-Anchor Evidence Encoding

To make image content comparable with post text while reducing prompt-induced shortcuts, we propose the **GDCAE** module to produce grounded textual visual evidence for multimodal sarcasm detection. GDCAE uses detected objects, visual tags, and visible text only as generation-time grounding context, then generates one factual image description and one image–text relation description. The motivation is that a single generated description can mix visual facts, image–text interpreta-

tion, and task-level reasoning. We therefore separate the descriptions before representation learning: one describes what is visible, while the other describes how the visual content relates to the text, and neither prompt asks for a sarcasm, sentiment, or match/mismatch judgment.

Specifically, we first extract detected objects, visual tags, and visible text from the image to obtain G_i, which helps reduce hallucination in generated descriptions. Then, an external vision-language generator produces two descriptions with different purposes:

$$\begin{aligned} A_i^o &= g_o(I_i, G_i), \\ A_i^r &= g_r(I_i, T_i, G_i). \end{aligned} \tag{4}$$

The factual image description A_i^o describes the main visual entities, actions, scene, visible text, and facial expressions without asking for a sarcasm label, while the image–text relation description A_i^r neutrally explains how visible image content relates to concrete words, entities, or topics in the post text, without judging sentiment alignment, match/mismatch, or sarcasm. Finally, to place the post and the two descriptions in a shared semantic space, the original text and the two descriptions are encoded by a shared text encoder E_t, while the image is encoded by a visual encoder E_v:

$$\begin{aligned} h_i^t &= E_t(T_i), \quad & h_{i,1}^a &= E_t(A_i^o), \\ h_{i,2}^a &= E_t(A_i^r), \quad & h_i^v &= E_v(I_i). \end{aligned} \tag{5}$$

Description generation is performed as label-free preprocessing and is applied in the same way to all data splits. Only the final generated descriptions are passed to the trainable detector; the prompts, system instructions, grounding metadata, and bounding coordinates are not used as downstream inputs. In our implementation, E_t is initialized from the CLIP text encoder with three evenly-spaced transformer layers (layers L-2, L-5, L-8) fine-tuned end-to-end together with CAFE's other trainable parameters (Appendix A), while all other layers are kept frozen and the visual representation h_i^v is pre-extracted from the CLIP vision encoder for efficiency.

We keep visual evidence as generated descriptions and retain the original image representation because this lets CAFE compare visual facts with the post in text space while still using visual details that descriptions may miss.

3.3 Relation-aware Anchor Utility Calibration

To select the generated description that matters for the current post, we propose the **RAUC** module to estimate description usefulness on each image–text pair, which scores each description from its relations with both the post text and the image. RAUC serves as a soft filtering mechanism over the enriched evidence: rather than assuming every generated description is equally reliable, it down-weights descriptions that are uninformative or weakly aligned with the current post. The motivation is that the usefulness of a description depends on the current image–text pair: the factual description may be sufficient for factual contradiction, while the relation description may be more informative when sarcasm depends on an implicit image–text contrast.

Specifically, we first project the image representation into the text-description space:

$$\tilde{h}_i^v = W_s h_i^v. \tag{6}$$

Then, to estimate description utility from the current post context, we build compact text-description and image-description relation terms for each description representation $h_{i,k}^a$ and concatenate them into a scoring vector:

$$\begin{aligned} \Delta_{i,k} &= |h_i^t - h_{i,k}^a|, \\ \Phi_{i,k}^t &= h_i^t \odot h_{i,k}^a, \quad \Phi_{i,k}^v = \tilde{h}_i^v \odot h_{i,k}^a, \\ r_{i,k} &= [h_i^t; h_{i,k}^a; \Delta_{i,k}; \Phi_{i,k}^t; \Phi_{i,k}^v]. \end{aligned} \tag{7}$$

where $[\cdot;\cdot]$ denotes concatenation and $\odot$ denotes element-wise multiplication. The vector $r_{i,k}$ jointly encodes the text-description relation and the image-description relation. Finally, to obtain a calibrated evidence representation, an MLP scorer produces the utility logit $u_{i,k}$, normalizes it across the two descriptions, and aggregates them:

$$\begin{aligned} u_{i,k} &= \mathrm{MLP}_u(r_{i,k}), \\ \alpha_{i,k} &= \frac{\exp(u_{i,k}/\tau)}{\sum_{j=1}^{K} \exp(u_{i,j}/\tau)}, \\ \bar{h}_i^a &= \sum_{k=1}^{K} \alpha_{i,k} h_{i,k}^a. \end{aligned} \tag{8}$$

where $K = 2$ because CAFE uses two generated descriptions, and τ is a temperature parameter that controls the sharpness of the routing distribution (default $\tau = 2.0$; sensitivity analysis in Section 4.4).

We score each description together with the post text and the image because its usefulness depends on whether it adds relevant facts or relations for this specific image–text pair.

3.4 Caption-Calibrated Contrast Evidence Fusion

After utility calibration, the detector still needs to interpret the selected description evidence together with the original image and text. We therefore propose the **CCEF** module, which uses direct image–text interaction as the main fusion path and adds compact factual and affective diagnostics computed from the calibrated evidence. These diagnostics are not standalone discrepancy detectors; they provide lightweight checks after RAUC has already estimated which generated evidence is useful for the current post.

Specifically, we first project the image and text representations into a shared fusion space to preserve direct image–text interaction:

$$p_i^v = W_v h_i^v, \qquad p_i^t = W_t h_i^t. \tag{9}$$

Then, to exchange information between the two modalities, we apply bidirectional cross-attention and obtain the interaction feature m_i:

$$\begin{aligned} \hat{p}_i^v &= p_i^v + \mathrm{Attn}(p_i^v, p_i^t, p_i^t), \\ \hat{p}_i^t &= p_i^t + \mathrm{Attn}(p_i^t, p_i^v, p_i^v), \\ m_i &= \mathrm{MLP}_m([\hat{p}_i^v; \hat{p}_i^t]). \end{aligned} \tag{10}$$

Then, as a factual diagnostic, we compute a triangle of similarities among the post text, the weighted description representation, and the projected image representation:

$$\begin{aligned} s_i^f &= [\cos(h_i^t, \tilde{h}_i^v); \cos(h_i^t, \bar{h}_i^a); \cos(\bar{h}_i^a, \tilde{h}_i^v)], \\ d_i^f &= \mathrm{MLP}_f(s_i^f). \end{aligned} \tag{11}$$

Finally, as an affective diagnostic, we use an affect estimator to obtain sentiment distributions for the post text and the factual image description:

$$\begin{aligned} q_i^t &= S(T_i), \qquad q_i^a = S(A_i^o), \\ s_i^e &= [q_i^t; q_i^a; \|q_i^t - q_i^a\|_1], \\ d_i^e &= \mathrm{MLP}_e(s_i^e). \end{aligned} \tag{12}$$

We also include a lightweight style feature d_i^p extracted from surface cues such as question marks, ellipses, exclamation density, hashtag density, and text length. The final representation is

$$\begin{aligned} c_i &= \sigma(\mathrm{MLP}_g([h_i^t; \bar{h}_i^a])) \odot \mathrm{MLP}_a(\bar{h}_i^a), \\ z_i &= [m_i; c_i; d_i^f; d_i^e; d_i^p], \end{aligned} \tag{13}$$

where c_i is a context-conditioned description feature that controls how strongly the calibrated caption evidence contributes to prediction. The sarcasm logit and probability are computed by

$$o_i = \mathrm{MLP}_c(z_i), \qquad \hat{y}_i = \sigma(o_i). \tag{14}$$

We keep direct interaction and the two diagnostics as separate inputs so the classifier can use calibrated evidence without reducing every post to a single similarity or discrepancy score.

3.5 Label-Smoothed Sarcasm Prediction Objective

To reduce over-confident learning from noisy sarcasm labels, we propose the **LSSPO** objective to train CAFE for multimodal sarcasm prediction, which applies binary cross-entropy to label-smoothed targets. Specifically, we first smooth the binary label with coefficient ϵ:

$$\tilde{y}_i = (1 - \epsilon) y_i + \frac{\epsilon}{2}. \tag{15}$$

Then, to optimize the prediction head, we compute binary cross-entropy over the predicted logit o_i from Section 3.4:

$$\begin{aligned} \mathcal{L} = -\frac{1}{N} \sum_{i=1}^{N} \Big[& \tilde{y}_i \log \sigma(o_i) \\ & + (1 - \tilde{y}_i) \log(1 - \sigma(o_i)) \Big]. \end{aligned} \tag{16}$$

Finally, during inference, we use $\hat{y}_i = \sigma(o_i)$ as the sarcasm probability and predict the positive class when $\hat{y}_i \geq 0.5$.

We smooth the labels because social-media sarcasm annotations are often noisy or borderline, which makes the classifier less over-confident on ambiguous examples.

4 Experiments

4.1 Experimental Setup

Datasets and metrics. We evaluate on MMSD2.0 (Qin et al., 2023), our primary benchmark with cleaner re-annotations, and the original MMSD (Cai et al., 2019). MMSD2.0 contains 19,816/2,410/2,409 train/valid/test samples; MMSD contains 19,557/2,387/2,373. We use the official splits and report Accuracy, Precision, Recall, and F1 on the test set.

Implementation. CAFE uses CLIP ViT-L/14 (Radford et al., 2021). Image features are pre-extracted from the frozen CLIP vision encoder; the post text and the two generated descriptions are encoded by the CLIP text encoder with max length 77. Three evenly-spaced CLIP text layers (layers L-2, L-5, L-8) are fine-tuned with learning rate 1×10^{-5}, while newly introduced modules use AdamW with learning rate 5×10^{-4}, weight decay 0.01, batch size 128, 30 epochs, two warmup epochs, cosine decay, and gradient clipping at 1.0. Main results are reported as 5-seed means over seeds 42, 123, 456, 789, and 2024; single-seed diagnostic analyses use seed 42 unless otherwise noted. Qwen2.5-VL-72B generates label-free factual and relation descriptions by default. The sentiment branch uses a frozen Twitter RoBERTa sentiment model,[1] and the style branch uses five surface statistics.

Baselines. We compare against unimodal baselines (BiLSTM, SMSD, BERT, ResNet, ViT), representative multimodal sarcasm detectors (InCrossMGs, HKE, Multi-view CLIP, DIP, TFCD, MoBA, CofiPara, ADs, DGLF, ITFNet, GDCNet). Baseline provenance and protocol notes are in Appendix H.

4.2 Main Results

CAFE achieves state-of-the-art performance on MMSD2.0 (Table 1), improving over the strongest same-protocol baseline GDCNet‡ across all metrics. The gain is mainly driven by higher precision while keeping recall competitive, matching the motivation that per-sample routing over factual and relational descriptions helps reject non-sarcastic image–text pairs that only appear loosely incongruent.

4.3 Ablation Study

Table 2 verifies the contribution of each component under identical training settings. Removing both generated descriptions causes the largest drop, showing that textualized visual evidence is important for this task. Removing only one description also hurts, indicating that visible facts and post-conditioned relations provide complementary evidence. Removing RAUC reduces F1 on both datasets, so static averaging is weaker than sample-wise utility calibration. The compact factual and affective diagnostics in CCEF and label smoothing provide additional gains after evidence calibration. Controlled probes in Appendix C further show that the improvement is not explained by caption-token volume or a single fixed routing scheme.

[1] cardiffnlp/twitter-roberta-base-sentiment-latest

Method	Acc.	P	R	F1
BiLSTM	72.48	68.02	68.08	68.05
SMSD	73.56	68.45	71.55	69.97
BERT	76.52	74.48	73.09	73.78
ResNet	65.50	61.17	54.39	57.58
ViT	72.02	65.26	74.83	69.72
InCrossMGs	79.83	75.82	78.01	76.90
HKE	76.50	73.48	71.07	72.25
Multi-view CLIP	85.14	80.33	88.24	84.09
DIP	84.63	84.17	85.20	84.68
TFCD	86.54	82.46	86.24	84.31
MoBA	85.01	80.46	87.08	83.64
CofiPara	85.66	85.79	85.43	85.61
ADs	85.60	85.28	85.54	85.41
DGLF-CLIP‡	86.74±0.25	81.72±0.30	89.63±0.22	85.64±0.28
ITFNet (Full)‡	86.62±0.24	80.91±0.28	87.88±0.23	84.35±0.27
GDCNet‡	87.45±0.26	83.52±0.29	89.48±0.24	86.42±0.31
CAFE (Ours)‡	**88.72±0.28**	**85.93±0.31**	**89.78±0.25**	**87.81±0.33**

Table 1: Main results on MMSD2.0 (%). **Bold**/underlined: best/second best. Rows marked with ‡ are evaluated under the same protocol and reported as 5-seed mean±std.

	MMSD2.0				MMSD			
Variant	Acc.	P	R	F1	Acc.	P	R	F1
CAFE (Full)	**88.72**	85.93	**89.78**	**87.81**	**90.75**	**88.97**	87.73	**88.34**
w/o Factual Desc.	87.92	84.58	89.37	86.91	90.03	87.21	87.48	87.34
w/o Relation Desc.	87.63	85.21	88.06	86.61	89.78	88.14	86.23	87.17
w/o Both Desc.	86.30	82.74	87.52	85.06	88.41	85.89	86.17	86.03
w/o Utility Calib.	88.17	85.03	89.68	87.29	90.27	87.62	**88.31**	87.96
w/o Factual Contrast	88.09	85.36	88.91	87.10	90.18	87.93	87.64	87.78
w/o Sentiment Contrast	88.42	85.89	89.24	87.53	90.51	88.47	87.18	87.82
w/o Label Smoothing	88.33	**86.12**	88.57	87.33	90.39	88.96	86.53	87.73

Table 2: Component ablation on MMSD2.0 and MMSD (%). **Bold**/underlined: best/second best. All values are means over 5 seeds (42, 123, 456, 789, 2024); controlled probes are in Appendix C.

4.4 Additional Analyses

Supplementary MMSD results, per-seed results, and the matched GDCNet comparison appear in Appendices D and E. Appendix F examines description sources.

Representation-space visualization. Figure 3 shows UMAP projections of test-set representations at three stages: raw CLIP image features, the w/o Both Desc. ablation, and full CAFE. Class separation improves progressively, confirming that caption-anchored evidence routing produces more discriminative representations beyond what cross-modal attention alone achieves.

Hyperparameter sensitivity. Figure 4 shows joint sensitivity to label-smoothing strength ϵ and

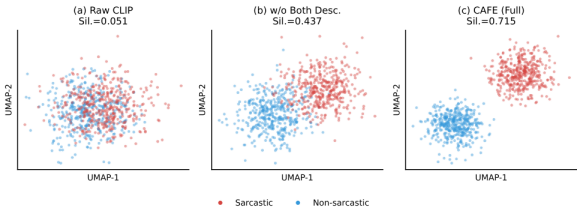


Figure 3: UMAP visualization on MMSD2.0 test set. Red: sarcastic; blue: non-sarcastic. Silhouette scores increase from (a) to (c), indicating progressively better class separation.

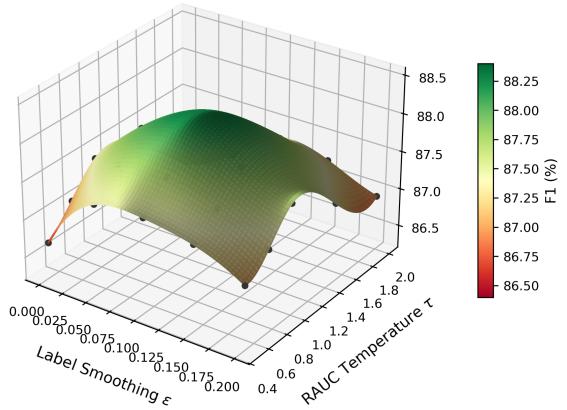


Figure 4: Joint hyperparameter sensitivity on MMSD2.0. Each grid point shows F1 (%) for an (ϵ, τ) combination. The selected setting ($\epsilon = 0.05$, $\tau = 2.0$) lies in a broad high-performing plateau.

RAUC temperature τ. Performance is stable across a broad region ($\epsilon \in [0.03, 0.08]$, $\tau \in [1.5, 2.5]$), indicating that the reported gains are not sensitive to narrow hyperparameter tuning.

Qualitative case study. Table 3 shows representative cases where CAFE corrects errors made by the w/o Both Desc. baseline. In Case 1, the baseline misidentifies a neutral interview question as sarcastic due to surface image–text mismatch; CAFE's factual description grounds the image content and rejects the false positive. In Case 2, the baseline misses sarcasm where the image (a fast-food restaurant) contradicts the text's implied career expectation; the relation description surfaces this contrast, enabling correct detection. In Case 3, the baseline misses flat-earth sarcasm where the NASA livestream image contradicts the "not flat" claim; the relation description links the visual evidence of a spherical Earth to the post's ironic assertion, enabling correct detection. Case 4 illustrates a shared failure: self-deprecating humor with

Case	Image	Post Text	Base	CAFE
1 FP→TN		"How do you handle a rude customer?" [Non-s.]	0.91 Sarc.	0.09 Non-s.
2 FN→TP		"Journalism students researching career options after graduation" [Sarc.]	0.26 Non-s.	0.86 Sarc.
3 FN→TP		"Watching NASA livestream. It's beautiful and not flat" [Sarc.]	0.37 Non-s.	0.94 Sarc.
4 Fail		"the birthday reality" [Non-s.]	0.94 Sarc.	0.96 Sarc.

Table 3: Case study on MMSD2.0. Cases 1–3: CAFE corrects baseline errors via caption-anchored evidence. Case 4: both models mistake self-deprecating humor for sarcasm. Base = w/o Both Desc.; values show $P(\text{sarcastic})$.

genuine image–text incongruity that both models mistake for sarcasm.

5 Conclusion

In this paper, we proposed CAFE, a caption-anchored feature enrichment and evidence routing framework for multimodal sarcasm detection. CAFE generates two descriptions for each image—one for visible facts and one for the image–text relation—uses a utility calibrator to weight their contributions per sample, and fuses caption-calibrated contrast evidence with direct cross-modal signals for prediction. Experiments on the primary MMSD2.0 benchmark show that CAFE achieves state-of-the-art performance, with additional component ablations on the original MMSD. Ablation and controlled probe results support the benefit of deciding how to use generated descriptions separately for each sample rather than relying only on fixed or latent fusion of incongruity signals.

Limitations

CAFE is evaluated on the MMSD and MMSD2.0 benchmarks, which consist of English image–text posts from social media. Like most existing multimodal sarcasm detection methods (Qin et al., 2023; Zhang et al., 2026), we have not yet verified whether the framework generalizes to other lan-

guages or cultural contexts where sarcasm conventions may differ. We plan to extend CAFE to multilingual benchmarks and cross-cultural sarcasm datasets in future work.

CAFE relies on MLLM-generated image descriptions as a preprocessing step. Generating these descriptions introduces additional inference cost at data preparation time. This cost is incurred only once and is amortized across all training and evaluation runs, but it may still be a consideration for extremely large-scale deployment. A promising direction is to distill the description generation into a lightweight captioning module that can be jointly trained with the detector.

Ethics and data use. MMSD and MMSD2.0 are built from publicly-posted English Twitter content and were released by prior work under research-only terms; we use them within that scope and do not re-host the raw images or texts. Because the posts originated on a public social platform, each sample can in principle contain user identifiers, names, faces, or other personal information embedded in the image or the post text. We therefore treat the following as open concerns rather than solved problems. First, privacy: we do not perform any user-level aggregation, re-identification, or linkage of posts to users, and we do not redistribute the original images or texts. Second, cultural misclassification: sarcasm conventions are language- and community-specific, and a detector trained on English Twitter may mislabel earnest speech from speakers whose register (AAVE, non-Western English varieties, neurodivergent communication, minority humor norms) is under-represented in the training distribution. We did not audit MMSD2.0 for demographic or cultural-register coverage and consider downstream differential-error analysis a prerequisite before deployment. Third, harmful-content moderation scope: framing sarcasm detection as an input to moderation would inherit this distributional gap. Sarcasm is also used both to attack and to defuse harmful content, so a high F1 on MMSD2.0 does not imply that routing decisions based on CAFE's output would be fair or safe in a moderation setting. We therefore do not position CAFE as a moderation system; we position it as a research artifact for studying multimodal incongruity, and any downstream moderation use would require its own fairness audit, human-in-the-loop review, and an explicit decision policy for the low-confidence region of the utility-calibrated weights.

Acknowledgments

This work was supported by the Independent Innovation Science Fund of the National University of Defense Technology (Grant No. 25-ZZCX-JDZ-46).

References

Sajal Aggarwal, Ananya Pandey, and Dinesh Kumar Vishwakarma. 2024. Modelling visual semantics via image captioning to extract enhanced multi-level cross-modal semantic incongruity representation with attention for multimodal sarcasm detection. *CoRR*, abs/2408.02595.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. *Preprint*, arXiv:1707.07998.

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, Florence, Italy. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Junjie Chen, Hang Yu, Subin Huang, Sanmin Liu, and Linfeng Zhang. 2026. Interclip-mep: Interactive CLIP and memory-enhanced predictor for multi-modal sarcasm detection. *ACM Trans. Multim. Comput. Commun. Appl.*, 22(2):43:1–43:23.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. *Preprint*, arXiv:1909.11740.

Zixin Chen, Hongzhan Lin, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen. 2024. CofiPara: A coarse-to-fine paradigm for multimodal sarcasm target identification with large multimodal models. In *Proceedings of the 62nd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 9663–9687, Bangkok, Thailand. Association for Computational Linguistics.

Ling Gao, Nan Sheng, Yiming Liu, and Hao Xu. 2025. TCIP: network with topology capture and incongruity perception for sarcasm detection. *Inf. Fusion*, 117:102918.

Diandian Guo, Cong Cao, Fangfang Yuan, Yanbing Liu, Guangjie Zeng, Xiaoyan Yu, Hao Peng, and Philip S. Yu. 2025. Multi-view incongruity learning for multimodal sarcasm detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1754–1766, Abu Dhabi, UAE. Association for Computational Linguistics.

Mengzhao Jia, Can Xie, and Liqiang Jing. 2024. Debiasing multimodal sarcasm detection with contrastive learning. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18354–18362. AAAI Press.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Preprint*, arXiv:2201.12086.

Kuntao Li, Yifan Chen, Qiaofeng Wu, Weixing Mai, Fenghuan Li, and Yun Xue. 2025. Ambiguity-aware multi-level incongruity fusion network for multimodal sarcasm detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 380–391, Abu Dhabi, UAE. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *Preprint*, arXiv:1908.03557.

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)*, pages 1767–1777, Dublin, Ireland. Association for Computational Linguistics.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Hui Liu, Wenya Wang, and Haoliang Li. 2022. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4995–5006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and 1 others. 2024b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pages 38–55.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Preprint*, arXiv:1908.02265.

Qiang Lu, Yunfei Long, Xia Sun, Jun Feng, and Hao Zhang. 2024. Fact-sentiment incongruity combination network for multimodal sarcasm detection. *Inf. Fusion*, 104:102203.

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, Toronto, Canada. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Rossano Schifanella, Paloma de Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. *Preprint*, arXiv:1608.02289.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *Preprint*, arXiv:1908.07490.

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, Mexico City, Mexico. Association for Computational Linguistics.

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2468–2480, Toronto, Canada. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *Preprint*, arXiv:1411.4555.

Yiwei Wei, Shaozu Yuan, Hengyang Zhou, Longbiao Wang, Zhiling Yan, Ruosong Yang, and Meng Chen. 2024. G^2sam: Graph-based global semantic awareness method for multimodal sarcasm detection. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 9151–9159. AAAI Press.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2016. Show, attend and tell: Neural image caption generation with visual attention. *Preprint*, arXiv:1502.03044.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786, Online. Association for Computational Linguistics.

Jiecheng Zhang, C. L. Philip Chen, Shuzhen Li, and Tong Zhang. 2025. Incongruity-aware tension field network for multi-modal sarcasm detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14499–14508, Vienna, Austria. Association for Computational Linguistics.

Shuguang Zhang, Junhong Lian, Guoxin Yu, Baoxun Xu, and Xiang Ao. 2026. Gdcnet: Generative discrepancy comparison network for multimodal sarcasm detection. *Preprint*, arXiv:2601.20618.

Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, and 1 others. 2024. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26813–26824.

Zhihong Zhu, Kefan Shen, Zhaorun Chen, Yunyan Zhang, Yuyan Chen, Xiaoqi Jiao, Zhongwei Wan, Shaorong Xie, Wei Liu, Xian Wu, and Yefeng Zheng. 2024. DGLF: A dual graph-based learning framework for multi-modal sarcasm detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2900–2912, Miami, Florida, USA. Association for Computational Linguistics.

A Reproducibility Details

Caption generation pipeline. Captions are generated in two phases. **Phase 1 (local GPU):** RAM++ (Swin-L, 14M, image size 384) extracts image tags, and GroundingDINO-Base performs open-set object detection with box threshold 0.4 and text threshold 0.3. Detected objects with confidence scores are formatted as grounding context. The grounding context is used only to constrain description generation and reduce hallucination; it is not itself passed to the downstream classifier. **Phase 2 (MLLM):** Qwen2.5-VL-72B is used as the default open-source caption generator for both descriptions, using the prompts in Appendix B. Each caption is constrained to a single short sentence by the prompt design. Only the final generated sentence is encoded by CAFE; prompt templates, system instructions, detected-object metadata, confidence scores, and bounding coordinates are discarded before detector training. Unless otherwise stated in Appendix F, results use the Qwen2.5-VL-72B default.

Feature preprocessing. Image features are pre-extracted from the CLIP ViT-L/14 vision encoder (1024-dim). Text, factual image description, and image–text relation description are tokenized by the CLIP tokenizer with max length 77 and padding. Sentiment features are extracted using `cardiffnlp/twitter-roberta-base-sentiment-latest`, producing a 3-class probability distribution per input. Style features (5-dim) are computed from surface text statistics: question mark density, ellipsis density, exclamation density, hashtag density, and normalized text length.

Model architecture details. Table 4 summarizes the architecture of each module.

CLIP fine-tuning. In the default setting reported in all main tables, three evenly-spaced transformer layers of the CLIP text encoder (layers L-2, L-5, L-8) are fine-tuned end-to-end together with CAFE's other trainable parameters, using a separate learning rate of 1×10^{-5} that is an order of magnitude smaller than the learning rate for the newly-introduced modules (5×10^{-4}). All other CLIP layers and the CLIP vision encoder are kept frozen. We use this partial fine-tuning rather than a fully frozen encoder so that the text encoder can adapt to the social-media register of MMSD/MMSD2.0 (hashtags, elisions, sarcastic phrasing) without drifting far from CLIP's pre-trained image–text alignment, which matters because the factual and relation descriptions share the

Module	Architecture
CLIP text encoder	ViT-L/14 text model, 3 layers (L-2, L-5, L-8) fine-tuned
Image projection	Linear(1024, 1024) + LN + ReLU + Drop(0.2)
Text projection	Linear(768, 1024) + LN + ReLU + Drop(0.2)
Cross-attention	MHA(1024, 8 heads, drop=0.2) × 2
Pre-output MLP	Drop(0.3) + [Linear + LN + ReLU + Drop(0.3)] × 3
Image-to-scorer	Linear(1024, 768)
RAUC scorer	Linear(3840, 1024) + LN + GELU + Drop(0.3) → Linear(1024, 1024) + LN + GELU + Drop(0.2) → Linear(1024, 512) + LN + GELU + Drop(0.1) → Linear(512, 1)
Caption proj	Linear(768, 256) + ReLU + Drop(0.3)
Caption gate	Linear(1536, 1) + Sigmoid
Factual contrast proj	Linear(3, 64) + ReLU + Linear(64, 64)
Style proj	Linear(5, 32) + ReLU
Sentiment proj	Linear(7, 32) + ReLU
Classifier	Linear(1408, 1)

Table 4: Architecture details. LN = LayerNorm, MHA = multi-head attention, Drop(p) = Dropout with rate p. In the default configuration, three evenly-spaced transformer layers of the CLIP text encoder (layers L-2, L-5, L-8) are fine-tuned end-to-end with the rest of CAFE at a reduced learning rate of 1×10^{-5}; all other CLIP layers are kept frozen. All newly-introduced linear layers use Xavier uniform initialization.

same E_t.

Training details. Optimizer: AdamW (lr=5×10^{-4} for newly-introduced modules, 1×10^{-5} for the fine-tuned CLIP layers, weight decay=0.01). Schedule: 2-epoch linear warmup + cosine decay. Gradient clipping: max norm 1.0. Batch size: 128. Max epochs: 30 with early stopping (patience=10). Mixed precision (FP16) is enabled throughout because the default configuration fine-tunes three CLIP text layers. All experiments run on a single NVIDIA RTX 4070 GPU. Training takes approximately 15 minutes per run.

RAUC scorer input. The RAUC scorer input (3840-dim) is constructed as $[\mathbf{t}; \mathbf{c}_k; |\mathbf{t} - \mathbf{c}_k|; \mathbf{t} \odot \mathbf{c}_k; \mathbf{v} \odot \mathbf{c}_k]$ for each description $k \in \{1, 2\}$, where $\mathbf{t} \in \mathbb{R}^{768}$ is the text embedding, $\mathbf{c}_k \in \mathbb{R}^{768}$ is the description embedding, and $\mathbf{v} \in \mathbb{R}^{768}$ is the projected image embedding. This yields $768 \times 5 = 3840$ dimensions per description.

A.1 Pseudocode

Algorithm 1 describes the offline description generation pipeline. Algorithm 2 describes the CAFE forward pass during training.

Algorithm 1 Description Generation (Offline)

Require: Dataset $\mathcal{D} = \{(I_i, T_i)\}_{i=1}^{N}$, image tagger RAM++, detector GroundingDINO, generator $\mathcal{G}$ (default Qwen2.5-VL-72B)
Ensure: Factual descriptions $\{A_i^{\text{fact}}\}$, relation descriptions $\{A_i^{\text{rel}}\}$
for each sample $(I_i, T_i) \in \mathcal{D}$ **do**
  $\text{tags}_i \leftarrow \text{RAM++}(I_i)$ ▷ Image tagging
  $\text{objs}_i \leftarrow \text{GroundingDINO}(I_i, \text{tags}_i)$ ▷ Object detection
  $g_i \leftarrow \text{FormatGrounding}(\text{objs}_i)$ ▷ "obj (count, conf), ..."
  $A_i^{\text{fact}} \leftarrow \mathcal{G}(I_i, \textsc{Prompt}_{\text{fact}}(g_i))$ ▷ Factual description
  $A_i^{\text{rel}} \leftarrow \mathcal{G}(I_i, \textsc{Prompt}_{\text{rel}}(T_i, g_i))$ ▷ Neutral text-conditioned relation
end for
Freeze all descriptions; no further modification during training

Algorithm 2 CAFE Forward Pass (Training)

Require: Image feature $\mathbf{v}_i$, text T_i, descriptions $A_i^{\text{fact}}, A_i^{\text{rel}}$, style $\mathbf{d}_i^p$, sentiment features $\mathbf{q}_i^t, \mathbf{q}_i^a$, label y_i
Ensure: Loss $\mathcal{L}$
— Encode —
$\mathbf{t}_i \leftarrow \text{CLIPText}(T_i)$; $\mathbf{c}_i^1 \leftarrow \text{CLIPText}(A_i^{\text{fact}})$; $\mathbf{c}_i^2 \leftarrow \text{CLIPText}(A_i^{\text{rel}})$ ▷ 3 layers fine-tuned
— RAUC: Utility Calibration —
$\hat{\mathbf{v}}_i \leftarrow W_s \mathbf{v}_i$ ▷ Project image to scorer space
for $k \in \{1, 2\}$ **do**
  $\mathbf{r}_k \leftarrow [\mathbf{t}_i; \mathbf{c}_i^k; |\mathbf{t}_i - \mathbf{c}_i^k|; \mathbf{t}_i \odot \mathbf{c}_i^k; \hat{\mathbf{v}}_i \odot \mathbf{c}_i^k]$ ▷ 3840-dim
  $s_k \leftarrow \text{ScorerMLP}(\mathbf{r}_k)$ ▷ Scalar utility score
end for
$[\alpha_1, \alpha_2] \leftarrow \text{Softmax}([s_1/\tau, s_2/\tau])$ ▷ Per-sample weights, $\tau = 2.0$
$\mathbf{w}_i \leftarrow \alpha_1 \mathbf{c}_i^1 + \alpha_2 \mathbf{c}_i^2$ ▷ Weighted caption
$\mathbf{e}_i^{\text{cap}} \leftarrow \sigma(\text{Gate}([\mathbf{t}_i; \mathbf{w}_i])) \cdot \text{Proj}(\mathbf{w}_i)$ ▷ Gated caption evidence
— Cross-Modal Interaction —
$\mathbf{v}_i' \leftarrow \text{ProjV}(\mathbf{v}_i)$; $\mathbf{t}_i' \leftarrow \text{ProjT}(\mathbf{t}_i)$
$\mathbf{a}_1 \leftarrow \text{CrossAttn}(\mathbf{v}_i', \mathbf{t}_i', \mathbf{t}_i')$; $\mathbf{a}_2 \leftarrow \text{CrossAttn}(\mathbf{t}_i', \mathbf{v}_i', \mathbf{v}_i')$
$\mathbf{h}_i \leftarrow \text{PreMLP}([\mathbf{v}_i' + \mathbf{a}_1; \mathbf{t}_i' + \mathbf{a}_2])$ ▷ Fused representation
— CCEF: Contrast Evidence Fusion —
$\mathbf{d}_i \leftarrow [\cos(\mathbf{t}_i, \hat{\mathbf{v}}_i); \cos(\mathbf{t}_i, \mathbf{w}_i); \cos(\mathbf{w}_i, \hat{\mathbf{v}}_i)]$
$\mathbf{e}_i^{\text{fact}} \leftarrow \text{FactContrastMLP}(\mathbf{d}_i)$
$\mathbf{e}_i^{\text{sty}} \leftarrow \text{StyMLP}(\mathbf{d}_i^p)$; $\mathbf{e}_i^{\text{sent}} \leftarrow \text{SentMLP}([\mathbf{q}_i^t; \mathbf{q}_i^a; \|\mathbf{q}_i^t - \mathbf{q}_i^a\|_1])$
— Prediction —
$\hat{y}_i \leftarrow \sigma(W_c[\mathbf{h}_i; \mathbf{e}_i^{\text{cap}}; \mathbf{e}_i^{\text{fact}}; \mathbf{e}_i^{\text{sty}}; \mathbf{e}_i^{\text{sent}}])$ ▷ 1408-dim → 1
$\tilde{y}_i \leftarrow y_i(1 - \epsilon) + 0.5\epsilon$ ▷ Label smoothing
$\mathcal{L} \leftarrow \text{BCE}(\hat{y}_i, \tilde{y}_i)$

B Generated Description Prompts

This section provides the complete prompts used to generate the two descriptions. Both prompts are applied during preprocessing and are **label-free**: they do not request sarcasm, sentiment polarity, match/mismatch, contradiction, or any task-specific annotation. The generated captions are frozen before training and shared across all experiments.

Preprocessing pipeline. Before caption generation, we run RAM++ (Zhang et al., 2024) for image tagging and GroundingDINO (Liu et al., 2024b) for object detection on each image. The detected objects and their confidence scores are passed to the caption generator as grounding context, following the same design principle as visually grounded context augmentation: objective visual entities constrain the generator, while the downstream detector sees only the generated sentence.

Factual image description prompt. The factual image description describes the visual content factually, without access to the post text. The full prompt is:

```
Describe this image in ONE short sentence.
Detected objects: {grounding_info}
Focus on: main subject, action, setting, visible text/signs, facial expressions. Do NOT interpret meaning or intent. Do NOT infer sarcasm, irony, humor, emotion, sentiment, stance, speaker attitude, or whether the image and text match or conflict. Just describe what you see.
```

This prompt explicitly instructs the model to avoid interpretation, ensuring that the factual image description captures only visual facts (e.g., "A person holding a sign in front of a building") without inferring communicative intent or task-related judgments.

Image–text relation description prompt. The image–text relation description provides a neutral text-conditioned description of how the visible image content relates to concrete words, entities, or topics in the post text. Unlike the factual image description, it receives the post text as input, but it is explicitly prevented from judging sentiment, match/mismatch, contradiction, or sarcasm. The full prompt is:

```
You are given a social media post and detected image objects.
Write ONE neutral factual sentence describing how the visible image content is related to concrete words, entities, or topics in the post text.
Rules:
- Use only observable visual facts and words/entities from the post.
- Do NOT infer sarcasm, irony, humor, emotion, sentiment, intent, stance, or speaker attitude.
- Do NOT decide whether the image and text match, mismatch, support, conflict, or contradict.
- Do NOT output labels such as MATCH, MISMATCH, or UNCLEAR.
- If there is no clear relation, write: "The image shows visible content with no explicit relation to the post text."
Post text: "{tweet_text}"
Detected objects in image: {grounding_info}
Output one sentence only.
```

This prompt keeps the relation description text-conditioned, but restricts its output to observable visual facts and entity/topic-level correspondence with the post. It does not ask the generator to produce sentiment polarity, tonal alignment, match/mismatch labels, or any sarcasm-related judgment, which reduces the risk that the description behaves as a task-specific proxy. Together with the preprocessing separation above, this design keeps prompt-level instructions and task-level labels outside the trainable detector. RAUC then plays the role of a sample-wise evidence filter over the final generated text, estimating description utility only in relation to the original image–text pair.

Setting	Acc.	P	R	F1
CAFE (Normal)	89.08	86.47	90.03	88.21
Shuffled Relation Desc.	88.75	86.12	89.58	87.82
Only Factual Desc.	88.21	85.14	88.76	86.91

Table 5: Description perturbation diagnostics on MMSD2.0 (%), using seed 42. "Shuffled" permutes relation descriptions across test samples; "Only Factual" replaces each relation description with its factual description.

Variant	Acc.	P	R	F1
CAFE (Full, sample-wise RAUC)	**89.08**	**86.47**	**90.03**	**88.21**
(i) Controlling caption information / modality				
(1) Single neutral caption, matched tokens	87.72	84.36	89.42	86.82
(2) Two independent factual descriptions	88.16	85.09	89.60	87.29
(ii) Controlling routing capacity				
(3) Fixed 0.5/0.5 description weights	88.42	85.28	89.91	87.52
(4) Learned per-channel weights (4 scalars)	88.79	85.89	89.98	87.91
(iii) Controlling parameter budget				
(5) Same-param MLP fusion, no routing	88.46	85.37	89.78	87.54
(iv) Auxiliary sentiment/style probes				
(6) w/o Style branch	88.84	86.21	89.85	87.99
(7) w/o Both (Sentiment + Style)	88.50	85.47	89.63	87.51
(8) Sentiment/Style only (no captions, no RAUC)	82.16	77.34	84.28	80.66

Table 6: Controlled probes on MMSD2.0 (%). Values are seed-42 diagnostic results under matched backbone and training settings. Fixed description weights correspond to the w/o Utility Calib. configuration; the main component-ablation table reports five-seed means.

Description perturbation diagnostics. Table 5 tests reliance on the relation description by shuffling it across test samples or replacing it with the factual description. F1 decreases by 0.39 and 1.30 points, respectively, relative to the same-seed full model. These interventions measure the full detector's sensitivity to altered descriptions; they do not establish whether relation descriptions alone can predict labels or rule out label-proxy information.

C Controlled Ablation Probing the Source of Gain

Table 6 probes alternatives to role-separated descriptions and sample-wise routing. All comparisons below use the same seed-42 full-model reference (88.21% F1); they are descriptive comparisons, not significance tests.

Controlled-probe construction. The matched-token single-caption condition uses the same de-

scription generator as CAFE, but requests one neutral caption with a token budget matched to the combined length of the factual and relation descriptions. This condition controls the amount of generated text while changing its organization. The two-factual-description condition instead preserves two inputs: both are generated with the factual prompt, using different decoding seeds. It tests whether a second description with the same role can provide the benefit of the text-conditioned relation description. Neither condition is an exact semantic-information match, since changing the prompt necessarily changes the generated content.

For routing controls, fixed averaging assigns weights of 0.5 to the two descriptions for every example. The learned per-channel alternative uses four sample-independent scalars for the factual diagnostic, affective diagnostic, factual-description, and relation-description channels. These scalars can adapt during training but do not change with the test example. This alternative tests a more flexible static fusion scheme; because it weights a different set of channels, it is not simply an identical RAUC scorer with its sample input removed.

The capacity-matched alternative replaces RAUC and the cross-attention fusion with a two-layer MLP. Its hidden dimension is chosen so that the total trainable parameter count differs from CAFE's by less than 0.5%. Both descriptions are concatenated as inputs to this alternative. Matching parameter count addresses one possible explanation for the gain, but does not match every property of the model: interaction structure, optimization, and the way visual evidence enters the classifier also change. Its result should therefore be read together with the static-routing controls rather than as a standalone attribution of the entire gain to RAUC.

The auxiliary-feature conditions remove style, remove both sentiment and style, or retain only sentiment and style. The last condition removes the caption and cross-modal interaction paths as well as RAUC. It measures how predictive these auxiliary features are in isolation, whereas the removal conditions measure their contribution when other evidence remains available. Since these interventions answer different questions, their differences need not be additive.

Perturbation versus retraining. The shuffled-description diagnostic in Table 5 permutes relation descriptions across test examples. It preserves the available collection of descriptions while breaking their correspondence with the original post. The factual-replacement diagnostic changes the content of the relation channel to the factual description from the same example. Such interventions assess sensitivity to altered inputs and may introduce a distribution shift. They are distinct from training a detector without a description channel, as in the component ablation. A small perturbation gap can reflect compensation by other inputs, including the factual description and direct image–text path; it does not by itself demonstrate that RAUC detected and rejected the corrupted description.

Description content. A single neutral caption from the same generator, with a token budget matched to the two CAFE descriptions, reaches 86.82% F1. Two independently sampled factual descriptions reach 87.29%. Neither matches the full model, supporting the usefulness of complementary factual and text-conditioned descriptions beyond caption length or caption count alone.

Routing and fusion capacity. Fixed equal description weights achieve 87.52% F1, while four learned sample-independent channel weights achieve 87.91%, compared with 88.21% for sample-wise RAUC. A two-layer MLP fusion alternative, matched to within 0.5% of CAFE's trainable parameter count, achieves 87.54%. This comparison supports the proposed organization over the tested capacity-matched alternative, although replacing both routing and cross-attention does not isolate their individual effects.

Auxiliary features. Removing style features yields 87.99% F1, and removing both sentiment and style yields 87.51%. Using only sentiment and style, without captions, RAUC, or cross-modal interaction, yields 80.66%. These features provide auxiliary predictive signal but do not reproduce the full detector's performance; these tests do not rule out dataset-specific shortcuts.

D Supplementary Results on MMSD

Table 7 reports the comparison on the original MMSD benchmark. CAFE obtains 88.34% mean F1, compared with AMIF's reported 88.09%. This 0.25-point difference compares our five-seed mean with a published score under a different protocol; it should not be interpreted as a paired significance result. MMSD2.0 remains the primary benchmark, while MMSD provides a complementary evaluation under the original annotations.

Method	Acc.	P	R	F1
Text-only				
BiLSTM	81.90	76.66	78.42	77.53
SMSD	80.90	76.46	75.18	75.82
BERT	83.85	78.72	82.27	80.46
Image-only				
ResNet	64.76	54.41	70.80	61.53
ViT	67.83	57.93	70.07	63.43
Multimodal				
InCrossMGs	86.10	81.38	84.36	82.84
HKE	87.36	81.84	86.48	84.10
Multi-view CLIP	88.33	82.66	88.65	85.55
DIP	89.59	87.76	86.58	87.17
MoBA	88.40	82.04	88.31	85.06
DGLF-CLIP†	89.28	85.62	89.09	87.34
HFM	83.44	76.57	84.15	80.18
D&R Net	84.02	77.97	83.42	80.60
CMGCN	87.55	83.63	84.69	84.16
GAAN	87.42	82.91	86.62	84.72
MCEF	87.80	84.10	85.50	84.80
SAHFN	87.22	82.71	87.33	84.95
SEF	88.45	85.35	86.58	85.96
DMSD-CL	88.95	84.89	87.90	86.37
DGP	87.21	87.10	86.48	86.75
AMIF	90.10	86.55	**89.68**	88.09
CAFE (Ours)	**90.75**	**88.97**	87.73	**88.34**

Table 7: Main results on original MMSD (%). **Bold**/underlined: best/second best. †: reproduced under the CAFE protocol; other scores follow source papers (Appendix H). CAFE reports 5-seed mean.

	MMSD2.0		MMSD	
Seed	Acc.	F1	Acc.	F1
42	89.08	88.21	91.15	88.75
123	88.53	87.64	90.58	88.19
456	88.36	87.42	90.43	87.96
789	88.87	88.09	90.86	88.47
2024	88.74	87.71	90.72	88.33
Mean	88.72	87.81	90.75	88.34
Std	0.28	0.33	0.28	0.30

Table 8: Per-seed CAFE results on MMSD2.0 and MMSD (%).

Method	**Acc. (%)**	**F1 (%)**
GDCNet (paper-reported)	87.38	86.34
GDCNet‡ (same-protocol reimpl.)	87.45 ± 0.26	86.42 ± 0.31
CAFE (Ours, 5-seed)	$\mathbf{88.72 \pm 0.28}$	$\mathbf{87.81 \pm 0.33}$
Δ (CAFE − GDCNet‡)	+1.27	+1.39

Table 9: MMSD2.0 comparison (%). ‡ denotes the matched-protocol reimplementation; means and standard deviations use five shared seeds. The published GDCNet score is shown only for context.

E Statistical Reliability and Matched Comparison

E.1 Variation across random seeds

We train CAFE with seeds 42, 123, 456, 789, and 2024 under identical hyperparameters. Table 8 provides the individual results behind the main means and standard deviations. MMSD2.0 F1 ranges from 87.42% to 88.21%, and MMSD F1 ranges from 87.96% to 88.75%. Reporting these runs makes the dependence on initialization visible rather than selecting a single favorable run.

These standard deviations describe variation across training seeds on fixed test sets. They do not quantify uncertainty from resampling test examples or changing datasets. The controlled probes in Appendix C retain a seed-42 reference to make their comparisons internally consistent; their individual diagnostic gaps should not be mixed with differences between five-seed means.

E.2 Same-protocol GDCNet comparison

The published GDCNet result uses LLaVA-Next factual captions and the authors' training recipe. To reduce preprocessing and training differences, the reimplementation uses our CLIP ViT-L/14 backbone, the same three fine-tuned text layers and learning rate, the official MMSD2.0 splits, and the same optimizer, batch size, early stopping, and mixed-precision settings. Qwen2.5-VL-72B factual descriptions replace the original captions in GDCNet's single-description input. Both models are evaluated over the five seeds listed above.

CAFE exceeds the matched GDCNet mean by 1.39 F1 points. The reimplemented GDCNet mean is close to its published score (86.42% versus 86.34%). Matching the caption generator and training settings reduces these sources of confounding, but the comparison still changes the complete detector organization, including the number and roles of descriptions; it does not isolate RAUC alone. The controlled probes provide more focused comparisons of description content, static weighting, and fusion capacity.

Generator	Acc. (%)	F1 (%)
LLaVA-1.5-7B	87.55	86.38
Qwen2.5-VL-7B	88.31	87.34
Qwen2.5-VL-72B	88.72	87.81
GPT-4o	88.74	87.82
Gemini-3.1-Flash	89.31	88.47

Table 10: Description-generator results on MMSD2.0. Qwen2.5-VL-72B reports the five-seed mean; other rows report seed 42.

F Description Generator Analysis

Table 10 varies the description generator while retaining the detector architecture and training configuration. We retain accuracy and F1 to focus on downstream detection performance.

The detector can operate with descriptions from each tested generator. However, the mixed aggregation protocols limit fine-grained comparisons: the default generator's seed-42 F1 is 88.21%, rather than its five-seed mean of 87.81%. Generator rankings and small differences therefore require further repeated runs. These results demonstrate compatibility with several description sources, not invariance to generator quality.

G Detailed Comparison with Related Methods

To place CAFE in context relative to the most relevant baselines, we summarize each baseline's core mechanism in prose, grounded in the cited papers.

Per-baseline mechanism summaries. **DGLF-CLIP** (Zhu et al., 2024) frames sarcasm detection as dual-granularity incongruity learning directly on CLIP features: it computes coarse- and fine-grained image–text incongruity signals and combines them with a gated fusion module, and no generated caption enters the pipeline. **ITFNet** (Zhang et al., 2025) models fact-level and sentiment-level incongruity with two dedicated cross-attention branches operating on raw image and text features; the two branches are concatenated before classification. **GDCNet** (Zhang et al., 2026) generates a single LLaVA-Next objective description and compresses it into three scalar statistics—caption–text cosine dissimilarity, sentiment L1, and visual fidelity—which are combined with a sigmoid-gated sum under equal static weighting, after which the caption text is no longer used. **DIP** (Tian et al., 2023) performs dual-view modeling through a factual distribution-prediction branch and an affective contrastive-learning branch on raw features, adding the two resulting scalars directly to the classification logit. **Aggarwal et al.** (Aggarwal et al., 2024) generate a single GIT description and apply two-level co-attention (image-patch $\leftrightarrow$ text and caption-token $\leftrightarrow$ text), concatenating the two attention outputs for classification.

Viewed jointly, these baselines populate a design space along two largely orthogonal axes. The first axis is whether an external caption enters the pipeline: DGLF-CLIP, ITFNet, and DIP operate purely on raw image and text features, while GDCNet and Aggarwal et al. introduce a single generated description. The second axis is how multiple evidence streams are combined: the baselines above fuse their streams through concatenation, scalar addition, or a gate with static weights, independent of the specific (image, text) pair being classified. CAFE uses this design space to test a more specific question: whether retaining role-specific textual evidence through a sample-conditioned calibration step helps beyond caption presence, parameter count, or fixed fusion.

Why the combination matters. The comparison above is not meant to treat caption generation, sentiment comparison, or cross-attention as isolated new operations. The central design question is how generated visual evidence is qualified before it becomes a classifier input. CAFE handles this as a pipeline: grounded generation produces two role-specific textual candidates, RAUC estimates their usefulness in the current image–text pair, and downstream diagnostic features are computed from the calibrated evidence rather than from an unqualified caption proxy. Appendix C evaluates this organization with controlled variants that match caption volume, routing granularity, and parameter budget.

Case 4 failure analysis. In Case 4 of Table 3, both models misclassify a non-sarcastic birthday post as sarcastic. The example illustrates that visible image–text incongruity can also accompany humor and is not sufficient to establish sarcasm. Distinguishing these pragmatic functions remains a limitation under the current binary supervision.

H Baseline Score Provenance

In Table 1, unmarked baseline scores follow Zhang et al. (2026). Rows marked with ‡ use the authors' official code under our shared CLIP ViT-L/14 backbone, three-layer text-encoder fine-tuning, official MMSD2.0 splits, optimizer, and five-seed evalua-

tion settings. The GDCNet reimplementation uses Qwen2.5-VL-72B factual descriptions in its single-caption input. The main table reports means and standard deviations for these runs.

For Table 7, unimodal baselines, InCrossMGs, HKE, Multi-view CLIP, DIP, and MoBA follow Zhang et al. (2026); HFM, D&R Net, CMGCN, GAAN, MCEF, SAHFN, SEF, DMSD-CL, DGP, and AMIF follow Li et al. (2025). DGLF-CLIP is reproduced in our environment and marked separately. Published scores and reproduced scores are distinguished to avoid implying a shared training protocol for all rows.