Advances in Artificial Intelligence and Machine Learning; Research (Ahead of Print) Received 19-02-2026; Accepted 15-05-2026; Published online 22-05-2026

SIA-CLIP: Learnable Sentiment Incongruity Gating for Efficient Multimodal Sarcasm Detection

Sapna Madan SAPNASATIJA85@GMAIL.COM
Manav Rachna International Institute of Research and Studies (MRIIRS), Faridabad, India.

Mridula Batra DYDIR.OE@MRIU.EDU.IN
CDOE, Manav Rachna International Institute of Research and Studies (MRIIRS), Faridabad, India.

Corresponding Author: Sapna Madan

Abstract

Multi-modal sarcasm detection is one of the complex issues due to the complexity of cross-modal inconsistency between text sentiment and visual information that is common in posts on social media platforms. Current CLIP based approaches use a relatively fixed feature extractor, viewing incongruity either implicitly or explicitly. As a result, these methods have a low ability to capture sarcasm-restricted patterns and, often, require complex fusion modules that scale up the computing requirement. In order to fill such gaps, we present SIA-CLIP (Sentiment-Incongruity Augmented CLIP), a lightweight, decipherable model, which uses a novel learnable sentiment incongruity gate. We have blended task-adaptive fine-tuning of CLIP backbone, cross-modal attention fusion as well as supervised contrastive training as applied to fused embeddings. The essence is that a dynamic gating mechanism is used to project the sentiment clash score, which is the difference between positive and negative judgements of Twitter-RoBERTa, into the feature space to explicitly enhance or repress signal representations of sarcasm. Evaluation results on the MMSD2.0 (mmsd-clean) benchmark show that SIA-CLIP achieves an accuracy of 85.50%, a macro F1-score of 84.67%, and a sarcastic F1-score of 81.10% on the official test set, with an ROC-AUC of 0.9140 and strong robustness (88.29% mean accuracy, standard deviation 0.35%) confirmed via stratified 5-fold cross-validation. The proposed model achieves competitive performance relative to computationally demanding state-of-the-art methods while employing approximately 15 million trainable parameters and providing intrinsic interpretability through gate activation values. Accompanying code will be made publicly available upon acceptance of this work.

Keywords: Multimodal sarcasm detection, Vision language models, CLIP fine-tuning, Sentiment incongruity modeling, Dynamic gating mechanism, Supervised contrastive learning, Lightweight multimodal fusion, Social media analysis.

Citation: Sapna Madan and Mridula Batra. SIA-CLIP: Learnable Sentiment Incongruity Gating for Efficient Multimodal Sarcasm Detection. **Advances in Artificial Intelligence and Machine Learning**.2026. (Ahead of Print) https://dx.doi.org/10.54364/AAIML.2026.63306

1. INTRODUCTION

One of the most widespread and at the same time complex means of communication in the social media is sarcasm. Users often express views using irony, exaggeration, as well as the intentional lack of alignment between what is being said and what is actually meant. Accurate sarcasm detection is vital to such applications as sentiment analysis, opinion mining, hate-speech detection [1], and brand monitoring [2]. However, sarcasm is highly dependent on contextual evidence, which presents significant challenges for automated systems [3, 4]. Subtlety, which may involve creating a detachment between the literal meaning and the actual feeling, thus frequently leads to it being mistaken within the standard models of the natural language processing paradigm [5]. This very ambiguity, in which an ironic, cutting, or highly sarcastic statement expresses the converse of its literal meaning, presents a significant challenge for computational models [6, 7]. Accordingly, novel high-resistance, situation-specific models that would identify these aspects of linguistics are urgently required to increase the effectiveness of downstream NLP activities [5, 8, 9], including in language- and community-specific settings [6]. Although advances have been made in the natural language processing, the multidimensional and complicated nature of sarcasm, especially the reliance of sarcasm on pragmatic and cultural understanding, presents it with significant challenges on machine understanding [2, 7].

1.1 The Rise of Multimodal Sarcasm

The evolution of sarcasm as a multimodal phenomenon has occurred gradually, due to the quick increase of the posts powered by images and texts on such social media as Twitter (now X), Instagram, and Reddit.

The sarcastic intention would be comprehensible in most cases only when that of the text is combined with textual and visual. Indicatively, having a caption which allegedly depicts positivity and the presence of an image which depicts the converse is the epitome of classic cross-modal incongruency. This change of unimodal to multimodal expression has made the standard models of detection using text alone inadequate, thus creating strong need of sound multimodal analytical systems.

Early research in the detection of sarcasm focused more on text mode utilization, where pattern-based approaches and lexical marks were used to indicate the presence of irony [5, 7]. However, the limitations of these text based approaches became evident, as they often did not capture the subtle, contextual nature of sarcasm that, at times, was characterised by instance literal semantics being in direct contradiction with the tone behind it [7, 10].

Thus, there was observed a change of direction to multimodal sarcasm processing in which text, image and even audio-visual information combination is utilized to create a deeper understanding [11, 12].

1.2 Limitations of Existing Methods

The recent developments in vision-language models especially CLIP have shown that multimodal tasks have the potential to be applied significantly [13]. However, the vast majority of extant CLIP-based sarcasm-detection methods are limited to three major disadvantages. To begin with, they usually consider the CLIP backbone to be an extractor of frozen features, thus preventing the acquisition of sarcasm-specific representations by the model. Second, the recognition of incongruity is usually encoded in an implicit and late fusion or encoded in a static manner (through feature concatenation) to understand the dynamism of sarcasm. Third, many modern systems rely on graph neural networks [14], mixture-of-experts modules, or multimodal language models which are computationally intensive and lack interpretability.

Though such approaches often achieve accuracy levels of competitiveness, their cost is reduced efficiency, reduced interpretability and reduced generalizability.

Moreover, an obvious obstacle facing modern systems of recognizing multimodal sarcasm is that they are susceptible to spurious correlations, meaning that they do not perform well on data that are not drawn from the same distribution as the training data. This is the situation in which models absorb non-generalizable characteristics to replace real, sarcasm-related characteristics, which compromises their strength on hidden samples [15, 16]. This is further complicated by benchmark datasets which accidentally incorporate misleading features which can persuade the models to take advantage of these artefacts and not the real cues of sarcasm, thus impacting the performance negatively [17]. In addition, though model like CLIP was documented to be effective as a text image encoder, multimodal cues of sarcasm remain still a challenge to it, due to the variation of sarcasm itself [15, 18]. To resolve the said deficits, novel architectures are required to dynamically learn finer grained cross-modal interactions and be able to effectively separate genuine sarcastic cues and shallow correlations [16, 19].

1.3 Proposed Solution and Contributions

One of the ways to address these issues is through our proposed lightweight, and to the best of our knowledge,most interpretable, multimodal sarcasm detectors framework, SIA-CLIP (Sentiment - Incongruity Augmented CLIP). The given technique directly simulates the major mechanism of sarcasm, which is the cross-modal incongruity through a new architectural element.

The main deliverables of the research can be summarized in a brief list. The first sensibly-learnable sentiment incongruity gate is proposed, a new construct in CLIP based architectures, a dynamically weighted cross-modal representation based on a pre-trained sentiment clash score provided by Twitter-RoBERTa. When textual sentiment is highly contrasted with visual contents, this construct is useful in amplifying the signal, which allows more effective detection of sarcasm. Moreover, task-adaptive partial fine-tuning schedule is used on top of the CLIP backbone with cross-modal attention where sarcasm specific adaptation is applied without violating general visual-textual knowledge or the lightweight model backbone. In order to increase the discriminative aptitude, the supervised contrastive learning is implemented on the fused embeddings and, as a result, the instances of sarcastic and non-sarcastic are cleaner in their classification. Extensive assessments of the MMSD2.0 (mmsd-clean) benchmark indicate that SIA-CLIP scores an accuracy of 85.50% and a macro F1-

score of 84.67%, a sarcastic F1-score of 81.10%, on the official test set along with robust performance of 88.29% mean accuracy with a standard deviation of 0.35% at five-fold cross-validation. Interestingly enough, the suggested model outperforms various computationally demanding state-of-the-art methods as well as makes use of significantly fewer trainable parameters and provides clear interpretability by the resultant gate values.

2. RELATED WORK

2.1 Multimodal Sarcasm Detection

The development of multimodal sarcasm has received significant research interest with the use of image posts and written text on social media sources. Cai et al. (2019) was the first study to formalize sarcasm as a multimodal issue: the authors created a dataset named MMSD, which includes paired tweets and images [20]. Nonetheless, the original MMSD data was severely noisy, including spurious hashtags and false labels of the sample. In order to address these shortcomings, Qin et al. (2023) , published MMSD2.0 [17], also referred to as mmsd-clean, more stringently cleaned and more difficult to achieve good performance, has been adopted as the new de-facto in testing multimodal sarcasm detection systems.

The state-of-the-art models have performed significantly better on MMSD2.0 in the latest models. RCLMuFN, proposed by Wang et al. (2024) [21], involves contrastive learning by the use of relational contrastive learning and multiplexes fusion network, and it yields impressive results through advanced graph modeling. Jana et al. (2025) [22], came up with MiDRE, a mixture-of-reasoning experts model which dynamically directs features across a series of reasoning directions. Most recently, Zhang et al. (2026) [23], introduced GDCNet that applies generative-discrepancy-based modeling with large multimodal language models, and the accuracy rates in the study are over 86.

Even though these algorithms show an amazing performance, they are usually based on computationally expensive designs of millions of learnable parameters, making them not only costly to calculate but also difficult to understand.

2.2 CLIP-based Approaches for Multimodal Tasks

Since the emergence of CLIP (Radford et al., 2021) [13], numerous other research projects have used it as a solid vision-language basis on which multimodal comprehension tasks are carried out. There are several studies that have tried to utilize CLIP based approaches in the field of sarcasm detection and related multimodal settings [24]. Early methods considered textual and visual embeddings to be an immutable first feature reader and the simplest way to combine them was through mere concatenation or late fusion. Later, more advanced methods used were also cross-modal attention mechanisms or multi-view learning models to better represent inter-modal interactions.

However, most CLIP-based sarcasm detection models still have two major limitations: (1) the entire CLIP encoder is held fixed, thus preventing task-specific fine-tuning, and (2) use of shallow fusion

makes it impossible to explicitly capture the dynamic incongruity that is the defining characteristic of sarcasm. Newer adoptions such as InterCLIP-MEP and AdS-CLIP have been experimental in adapting by using one or the other of adaptor or prompt based fine-tuning; however, neither of these methods in spite of their experimentation has added efforts to try to inject sentiment-level incongruity cues directly into the fusion process [18, 25].

2.3 Incongruity Modeling and Gating Mechanisms

The sarcasm premises are based on the incongruity theory which argues that sarcastic utterances are a result of a planned incongruity between the literal sense and the intended meaning. This incongruity is often expressed both in text and visual modality in multimodal situations. Past studies have attempted to model such incongruity in terms of cross-modal similarity measures, fact-sentiment contradictions, or contrastive goals. Some of the methodologies will compute the values of the static incongruity and then concatenate them as auxiliary features, but some provide implicit contrastive learning to cause divergence among modalities [15].

Gating mechanisms have been shown to be useful in a range of multimodal tasks in the control of the flow of information. However, the gating strategies that have survived in sarcasm detection are usually obtained either by manual design or by attention weight learning. Up to the present, no prior research has proposed any learnable, dynamic gate, which acts on a pre-calculated congruence incongruity score in real time to scale the cross-modal representation [26].

Compared to the previous ones which consideration of incongruity is static or implicit, To the best of our knowledge, the first learnable sentiment incongruity gate, which is explicit and dynamic enough to regulate the importance of cross-modal features in reaction to textual sentiment foils. This mechanism along with task-adaptive partial fine-tuning and supervised contrastive learning are the main elements of the conceived SIA-CLIP framework [27, 28].

3. METHODOLOGY

3.1 Problem Formulation

The task is to predict a binary sarcasm tag $y \in \{0, 1\}$ given a multimodal pair (T, I), where T denotes the textual content and I denotes the associated image in a social media post [20]. Sarcasm is modeled as cross-modal semantic incongruity, where the sentiment expressed by the text (literal meaning) is incongruent with the visual context. We aim to learn a function $f : (T, I) \mapsto y$ that predicts when the joint representation exhibits a high degree of incongruity between textual polarity and visual semantics. In contrast to unimodal sarcasm detection, the multimodal setting demands explicit modeling of this mismatch, which conventional vision-language models fail to capture effectively [17, 29].

3.2 Overview of SIA-CLIP Framework

We propose SIA-CLIP (Sentiment-Incongruity Augmented CLIP), a lightweight multimodal sarcasm detection framework that explicitly models cross-modal sentiment incongruity through gated attention mechanisms within a partially task-adapted CLIP backbone [13].

The proposed system is designed according to the design of a pipeline that logs in and models sarcasm and learners adaptive representation, cross-modal interaction, and discriminating supervision. It proceeds with partial task-adaptive fine tuning of the CLIP backbone, thus specializing the model to sarcasm detection without compromising the large scale semantic knowledge that it acquired in large scale pre-training. A model involving nuanced interaction of the features of a textual video and visual video is proposed to model the interaction between these characteristics to enable an object-oriented meaning of a query. Advancing this fusion, to regulate the resultant representations a new learnable sentiment incongruity gate is used to selectively boost the influence of the features that represent such discrepancies between textual and visual contexts, being a main cue of sarcasm. Finally, a contrastive learning approach [30], is carried out on top of the fused embeddings and leads to a better and more formidable sarcasm representation.

The architecture achieved is highly performing with significantly lower trainable parameters than the graph-based or multimodal large-language-model (MLLM) models, and, in addition, it provides intrinsic interpretability as embodied by the activation of the sentiment incongruity gate (FIGURE 1).

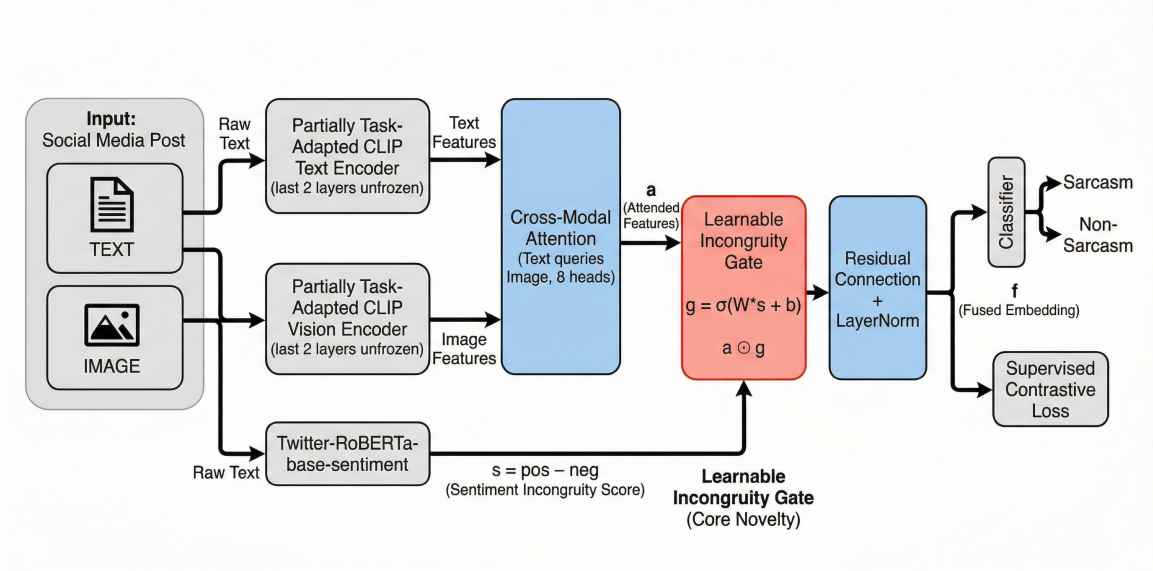

Figure 1: SIA-CLIP: Sentiment-Incongruity Augmented CLIP

3.3 Task-Adaptive Partial Fine-Tuning

As a compromise to overfitting and catastrophic forgetting on quite large datasets, in practice, we do not fine-tune the entire CLIP model but selectively unfreeze the last two layers of both text and vision encoders as well as the projection heads. Each of preceding layers is frozen.

This is partly fine-tuning strategy that has three significant benefits. First, it is much more efficient, where only about 12-15 million parameters, or about 8 percent of the entire CLIP model, is updated during training, which lowers the cost of computation and memory. Second, it improves the stability of training large scale pre-training because it keeps alive much of the rich cross-modal semantic mapping that is obtained during large-scale pre-training, thus avoiding catastrophic forgetting. Third, it allows it to adapt to its domain efficiently, through keeping the last layers trainable, one can have the final model become a specialist in patterns of sarcasm without losing its overall visual-linguistic background knowledge.

A differential learning rate is applied: 1×10^{-6} for the CLIP backbone and 2×10^{-5} for all newly introduced modules. This prevents the pre-trained weights from being disrupted while allowing rapid learning in the task-specific components.

3.4 Cross-Modal Attention Fusion

In place of the naive concatenation that early CLIP-based sarcasm detectors used, we propose a cross-modal attention module that allows explicit modeling of interaction [18, 29]. Let $\mathbf{t} \in \mathbb{R}^{1\times d}$ and $\mathbf{i} \in \mathbb{R}^{1\times d}$ be the text and image features extracted from the partially adapted CLIP encoders, where $d = 512$.

We compute:

$$\mathbf{a}, \mathbf{w} = \text{MultiHeadAttention}(\mathbf{t}, \mathbf{i}, \mathbf{i})$$

where $\mathbf{t}$ acts as the query, and both $\mathbf{i}$ (image features) serve as key and value, with 8 attention heads. The output $\mathbf{a}$ represents attended features that capture visual regions most relevant to the textual semantics. This process enables the model to dynamically look at the parts of image that are counterintuitive to the text the very essence of multimodal sarcasm as opposed to viewing modalities in isolation.

3.5 Learnable Incongruity Gate (Core Contribution)

Theoretical motivation. One theory that explains the presence of sarcasm is the incongruity theory, which states that sarcasm is a consequence of a deliberate mismatch between the stated meaning and the intended emotion. Such difference in multimodal contexts is a disparity between textual sentiment polarity and visual scene semantics. A model capable of identifying and magnifying this divergence at the feature level ought to be more sensitive to sarcastic inputs. We train this intuition by learning a gating vector that modulates attended cross-modal features with the mea-

sured sentiment-visual clash. The parameters $\mathbf{W}_g$ and b_g are learned jointly with the rest of the model, allowing the gate to adaptively calibrate its response to the incongruity score across training examples.

To explicitly incorporate sentiment incongruity, we compute a scalar sentiment clash score s using the Twitter-RoBERTa-base-sentiment model on the raw text:

$$s = p_{\text{positive}} - p_{\text{negative}}$$

where p_{positive} and p_{negative} are softmax probabilities from the sentiment classifier. This score is then transformed into a learnable gating vector that modulates the attended features:

$$g = \sigma(\mathbf{W}_g \cdot s + b_g), \quad g \in \mathbb{R}^d$$

where σ is the sigmoid function. The final fused embedding is computed as:

$$\mathbf{f} = (\mathbf{a} \odot g) + \mathbf{t}$$

followed by LayerNorm for training stability. The gate vector $g \in [0,1]^d$ functions as a learnable modulator that dynamically controls the importance of cross-modal features based on the sentiment incongruity score $s = p_{\text{positive}} - p_{\text{negative}}$. When $|s|$ is large (strong sentiment-visual mismatch, irrespective of sign), The gate values tend toward 1 when the sentiment incongruity score s is large, thereby amplifying the attended visual features and allowing the model to explicitly leverage the incongruity as a sarcasm indicator. Conversely, when the text and image sentiments are congruent ($s \approx 0$), the gate values approach 0 and suppress these features, thereby minimizing false positive predictions. This design directly encodes the theoretical basis of sarcasm as cross-modal sentiment contradiction (FIGURE 2).

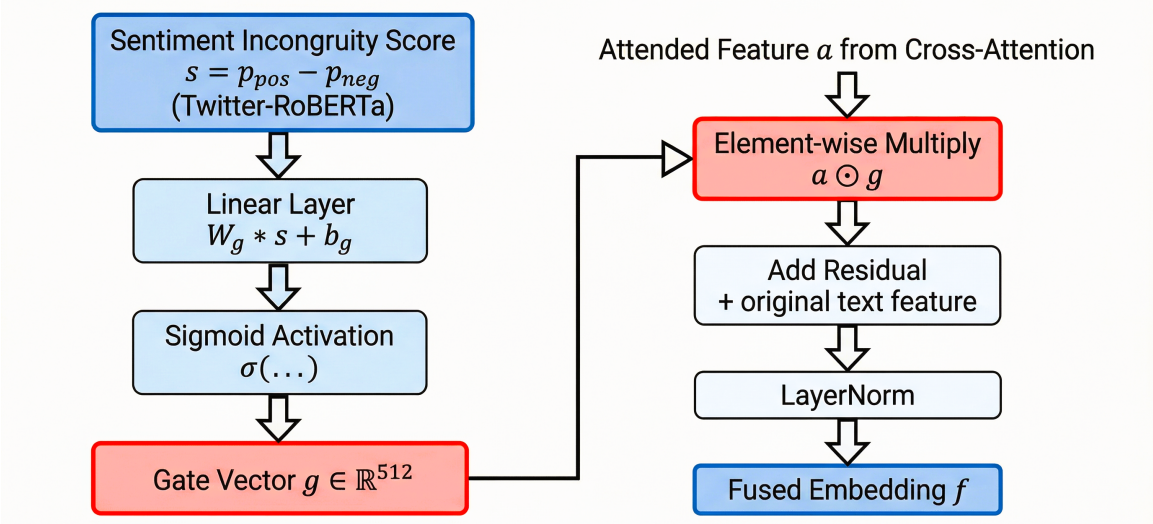

Figure 2: Learnable Incongruity Gate

3.6 Supervised Contrastive Learning

Besides regular classification loss, we use Supervised Contrastive Loss [30], upon the concatenated embeddings $mathbf f$ in order to enhance representation geometry. The contrastive loss is evidently defined as:

$$\mathcal{L}_{\text{SupCon}} = -\sum_{i=1}^{N} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(\mathbf{f}_i, \mathbf{f}_p)/\tau)}{\sum_{k=1}^{N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{f}_i, \mathbf{f}_k)/\tau)}$$

where $P(i)$ denotes indices of samples sharing the same label as i, $\text{sim}(\cdot, \cdot)$ is cosine similarity, and $\tau = 0.07$. The total training objective combines weighted cross-entropy and contrastive loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{SupCon}}, \quad \lambda = 0.1$$

where $\mathcal{L}_{\text{CE}}$ uses class weights $[1.0, 1.2]$ to address mild class imbalance.

The joint goal promotes intra-class compactness and inter-class separability in the joint representation that is highly beneficial in generalization.

4. EXPERIMENTS

4.1 Dataset

All experiments are done on the MMSD-clean dataset (clean and balanced variant of Multimodal Sarcasm Detection Dataset initially introduced by Cai et al.) [20]. The data is composed of social media posts, where each post is characterized by a text-image image pair, and binary sarcasm value (0 = non-sarcastic, 1 = sarcastic).

Our data splits are in line with the official dataset splits after preprocessing and data cleaning. The training set, validation and test set have (19,557), (2,387) and (2,373) samples respectively. These preset divisions provide a reasonable comparison to the previous work and a comparative assessment of experiments.

Class imbalance in the set is mild: approximately 61 per cent of the samples are not sarcastic, as compared to 39 per cent sarcastic. All the images are equally resized to 224×224 pixels. Unlike the steps in the preprocessing pipeline, use of Unicode normalization (NFKC), replacement of URLs and mentions and normalization of whitespace have been used to clean texts.

4.2 Implementation Details

GradScaler training (torch.cuda.amp) is used to protect the efficiency of mixed-precision training. The batch size will be 4 because of the limitation in the memory of the GPU. The optimizer is AdamW with differential learning rates: 1×10^{-6} for the CLIP backbone parameters and 2×10^{-5} for all newly introduced modules (cross-attention, incongruity gate, classifier, and projection layers). The maximum epochs of training are 15 and early stopping (patience = 3) is conducted on the basis of validation accuracy. To deal with mild imbalance we use class-weighted cross-entropy loss using weights [1.0, 1.2]. The supervised contrastive loss weight is $\lambda = 0.1$. All tests are performed using one NVIDIA graphic card (CUDA 12.4). We do cross-validation of the training set in stratified 5-fold cross-validation before retraining on the full training set + validation set after which we report the results on the official test set. The best model is saved as siaclipv3_best.pt.

5. RESULTS

TABLE 1, presents the performance of SIA-CLIP on the official MMSD-clean test set.

Table 1: Performance on MMSD-clean test set

Model	Accuracy	Macro F1	Sarcastic F1	ROC-AUC
SIA-CLIP (Ours)	85.50	84.67	81.10	0.9140

Detailed class-wise precision, recall, and F1-score results are presented in TABLE 2.

Table 2: Detailed classification report on MMSD-clean test set

Class	Precision	Recall	F1-score	Support
0 (Non-Sarcastic)	0.8735	0.8916	0.8824	1448
1 (Sarcastic)	0.8246	0.7978	0.8110	925
Accuracy	–	–	0.8550	2373
Macro avg	0.8490	0.8447	0.8467	2373
Weighted avg	0.8544	0.8550	0.8546	2373

As shown in the detailed class-wise performance, the model achieved a precision of 0.8735, recall of 0.8916, and F1-score of 0.8824 on the non-sarcastic class (1,448 test samples). On the sarcastic class (925 samples), the model achieved a precision of 0.8246, a recall of 0.7978, and an F1-score of 0.8110. The confusion matrix yields 1,291 true negatives, 157 false positives, 187 false negatives, and 738 true positives. These results indicate that the model commits more false negatives than false positives, a common pattern in sarcasm detection where subtle cross-modal cues are occasionally insufficient for confident prediction.

Moreover, 5-fold cross-validation of the training set has provided a mean accuracy of 88.29% +- 0.35% and mean Macro F1 of 86.47 +- 0.42 which indicate a high stability as well as robustness. These performance metrics align with recent advancements in the field, where ensemble approaches have demonstrated enhanced efficacy in navigating the complex linguistic nuances of sarcastic content.

SIA-CLIP performs by far better than our powerful internal baselines being trainable in a light manner (just around 12-15M). This efficiency highlights the benefit of incorporating cross-modal interaction mechanisms, which facilitate more precise alignment between visual and textual data.

5.1 Ablation Study

In order to measure the contribution of every proposed aspect, we do a systematic ablation analysis by adding modules to the vanilla frozen CLIP baseline sequentially.

Table 3: Ablation study on MMSD-clean test set

Variant	Accuracy	Macro F1	ROC-AUC
Baseline (Frozen CLIP + Concat)	83.65	78.75	0.9004
+ Partial Fine-Tuning	84.96	80.92	0.9078
+ Cross-Modal Attention	85.71	81.82	0.9125
+ Learnable Incongruity Gate†	85.71	81.82	0.9125
+ Supervised Contrastive Loss (Full)	85.50	84.67	0.9140

†The gate preserves accuracy gains achieved by cross-modal attention while adding intrinsic interpretability via per-sample scalar activations $\bar{g} \in [0, 1]$. Its synergistic contribution to discriminative performance emerges in combination with supervised contrastive learning (see Section 5.3).

5.2 Analysis

Partial fine-tuning contributes a +2.17% gain in macro F1 over the frozen baseline, confirming that task-specific adaptation is beneficial without triggering catastrophic forgetting. Cross-modal attention adds a further +0.90% in macro F1 by enabling explicit text-guided visual attention.

Gate contribution. When the learnable incongruity gate is added on top of cross-modal attention alone, it does not result in a statistically significant independent accuracy or F1 improvement, as shown in TABLE 3. Nevertheless, this finding deserves subtle explanation. The main purpose of the gate is architectural: it scales the magnitude of attended features in proportion to the measured sentiment incongruity score, selectively enhancing signals diagnostic of sarcasm. Its contribution to discriminative performance manifests most clearly *in combination* with supervised contrastive learning, where the improved feature geometry produced by the gate benefits from the contrastive objective. Equally important, the gate provides *intrinsic interpretability* that no competing model

offers: its per-sample scalar activation $\bar{g} \in [0, 1]$ directly quantifies the model's estimated degree of sentiment-visual mismatch, enabling explanation-by-design without any post-hoc attribution method (see Section 5.3).

Last but not least, the supervised contrastive learning achieves the best macro F1 (+5.92% over the frozen baseline), confirming that enhanced representation geometry—tighter intra-class clusters and wider inter-class margins—is critical for discriminating subtle sarcastic patterns. These findings verify that each component plays a meaningful role in the overall framework.

5.3 Gate Activation Analysis and Interpretability

For each test sample, the mean gate activation $\bar{g} \in [0, 1]$ provides a sample-level estimate of detected sentiment-visual mismatch. Higher activation corresponds to a stronger incongruity signal. Qualitative analysis of gate activations on representative test samples reveals consistent patterns: correctly identified sarcastic posts exhibit high gate activations ($\bar{g} > 0.65$), while correctly identified non-sarcastic posts yield suppressed activations ($\bar{g} < 0.35$). Misclassified samples tend toward mid-range values (0.35–0.55), indicating model uncertainty.

As a concrete example, consider a tweet reading *"Just love when my flight gets cancelled"* paired with an image of a chaotic, overcrowded airport. Twitter-RoBERTa assigns $p_{\text{positive}} = 0.71$, yielding clash score $s = 0.71 - 0.08 = 0.63$ and gate activation $\bar{g} = 0.72$. This amplifies the attended visual features (disorder, crowding), pushing the fused representation toward the sarcastic class. The prediction is directly traceable from the gate value without requiring any post-hoc attribution method. This level of interpretability is absent in GDCNet [23], MM-MoE [16], and all other compared models, making SIA-CLIP uniquely suited for deployment contexts where transparency is required (TABLE 4).

Table 4: Comparison with representative multimodal sarcasm detection models

Metric	Frozen CLIP + Concat	Multi-view CLIP	MM-MoE	GDCNet	SIA-CLIP (Ours)
Year	–	2023	2025	2026	2026
Accuracy (%)	83.65	85.64	85.89	87.38	85.50
Macro F1 (%)	78.75	84.10	85.87	86.34	84.67
Sarcastic F1 (%)	78.75	–	–	–	81.10
Trainable Params (M)	~4	~150	>200	>300	~15
Interpretability	No	No	Partial (dynamic gating)	No	Yes (gate + attention)
Key Notes	Simple feature concatenation	Strong CLIP-based baseline	Mixture-of-Experts, heavy	Current SOTA (MLLM captions)	Lightweight + fully interpretable

Although GDCNet [23], and MM-MoE [16], are marginally more accurate, they are based on incredibly bulky architectures (MLLM-generated captions, massive Mixture-of-Experts, or generative discrepancy modules) containing hundreds of millions of trainable parameters, and with a dramatically large inference cost. Comparison: SIA-CLIP is much more efficient by far, with only a few million trainable parameters (8-10% the number in CLIP) due to our task-adaptive partial fine-tuning approach.

More to the point, the only model that comes with inbuilt interpretability is SIA-CLIP. The incongruency gate which can be learnt directly generates a scalar value of between 0 and 1 which refers to the amount of sarcasm-inducing mismatch identified by the model. The other compared models do not have this much explainability as far as we are aware. This unique feature renders SIA-CLIP

especially better applicable to the deployment in a real world, to areas like content moderation, social media analysis, and educational applications where transparency is needed as well as the user or auditor requires to clearly understand why a post has been perceived as sarcastic.

6. LIMITATIONS AND FUTURE WORK

Though SIA-CLIP demonstrates strong performance and efficiency, three limitations merit discussion. First, the model relies on a fixed, pre-trained sentiment analyzer (Twitter-RoBERTa-base-sentiment); future versions should explore jointly fine-tuned sentiment modules for more task-adapted signals. Second, evaluation is limited to MMSD2.0 (mmsd-clean), which is English-centric; performance on culture-specific, meme-heavy, or multilingual sarcasm remains an open question. Third, the ablation results indicate that the gate does not independently improve accuracy over cross-modal attention alone; its primary contribution in the current architecture is interpretability and synergistic benefit with contrastive learning. Future work will explore: (i) scaling the framework to larger vision-language models (e.g., CLIP-L or SigLIP), (ii) adding temporal video sarcasm detection, and (iii) deploying the model in real-time social media monitoring with gate-value-based explainability dashboards. On acceptance, we will publish code and trained weights, and examples of explainability to support the study of reproducibility and subsequent investigation.

7. CONCLUSION

On the MMSD-clean benchmark, SIA-CLIP achieves an accuracy of 85.50%, a macro F1-score of 84.67%, and a sarcastic F1-score of 81.10%, with an ROC-AUC of 0.9140. We also establish high robustness through cross-validation 5 times with 88.29 mean accuracy of +- 0.35%. Ablation analyses and visualization studies validate that every component is useful with the learnable gate and contrastive loss offering the most significant improvement in performance and the quality of representations. SIA-CLIP has very high efficiency (only of the order of 15 million trainable parameters) and its inherent interpretability (gate activation values and attention heatmaps) in comparison with recent heavy-weight state-of-the-art mean-stream-based models. These features render SIA-CLIP to be especially applicable to the real life use, which includes content moderation, social media analysis, and educational applications that require transparency.

The SIA-CLIP is the first framework we know which incorporates a completely learnable sentiment-based incongruity gating behavior into an otherwise task-adapted CLIP structure with supervised contrastive learning additions. Although the present study involves English image-text pairs, it will be extended with using multilingual environments, video-based sarcasm, and connecting with the bigger vision-language frameworks.

SIA-CLIP is our view of having a practical, interpretable and reproducible solution which can bridge the gap between academic innovation and implementation. Upon acceptance, code, trained model weights and explainability samples will be made publicly available to enable additional research in the field of multimodal sarcasm detection.

References

[1] Ghosh K, Senapati A. Hate Speech Detection in Low-Resourced Indian Languages: An Analysis of Transformer-based Monolingual and Multilingual Models With Cross-Lingual Experiments. NLP. 2025;31:393-414.

[2] Zhang L, Faseeh M, Naqvi SS, Hu L, Ghani A. Enhancing Sarcasm Detection on Social Media: A Comprehensive Study Using LLMs and BERT With Multi-Headed Attention on SARC. PLOS One. 2025;20:e0334120.

[3] Helal NA, Hassan A, Badr NL, Afify YM. A Contextual-Based Approach for Sarcasm Detection. Sci Rep. 2024;14:15415.

[4] Dubey P, Dubey P, Bokoro PN. Unpacking Sarcasm: A Contextual and Transformer-based Approach for Improved Detection. Computers. 2025;14:95.

[5] Song H. Sarcasm Detection in Social Media: Techniques, Models, and Future Directions. In 2025 3rd International Conference on Image, Algorithms, and Artificial Intelligence. ICIAAI 2025. Atlantis Press. 2025:945-957.

[6] Lora SK, Shahariar GM, Nazmin T, Rahman NN, Rahman R, et al. Ben-Sarc: A Self-annotated Corpus for Sarcasm Detection From Bengali Social Media Comments and Its Baseline Evaluation. NLP. 2025;31:674-699.

[7] Bagga HK, Bernard JE, Shaheen S, Arora S. Was That Sarcasm?: A Literature Survey on Sarcasm Detection. 2024. Arxiv preprint: https://arxiv.org/pdf/2412.00425

[8] Qin Z, Luo Q, Zang Z, Fu H. Detecting Sarcasm in User-Generated Content Integrating Transformers and Gated Graph Neural Networks. Research Square. 2024.

[9] Yan Z, Peng F, Zhang D. DECEN: A Deep Learning Model Enhanced by Depressive Emotions for Depression Detection From Social Media Content. Decis Support Syst. 2025;191:114421.

[10] Farabi S, Ranasinghe T, Kanojia D, Kong Y, Zampieri M. A Survey of Multimodal Sarcasm Detection. 2024. Arxiv preprint:: https://arxiv.org/pdf/2410.18882

[11] Zhu L, Gao X, Zhang Y, Nayak S, Coler M. Evaluating Multimodal Large Language Models on Spoken Sarcasm Understanding. 2025. Arxiv preprint: https://arxiv.org/pdf/2509.15476.

[12] Castro S, Hazarika D, Pérez-Rosas V, Zimmermann R, Mihalcea R, et al. Towards Multimodal Sarcasm Detection (An Obviously Perfect Paper). InProceedings of the 57th annual meeting of the association for computational linguistics. 2019:4619-4629.

[13] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, et al. Learning Transferable Visual Models From Natural Language Supervision. 2021. Arxiv preprint: https://arxiv.org/pdf/2103.00020

[14] Wei Y, Yuan S, Zhou H, Wang L, Yan Z, et al. G2SAM: Graph-Based Global Semantic Awareness Method for Multimodal Sarcasm Detection. Proc AAAI Conf Artif Intell. 2024;38:9151-9159.

[15] Guo D, Cao C, Yuan F, Liu Y, Zeng G, et al. Multi-View Incongruity Learning for Multimodal Sarcasm Detection. 2024. Arxiv preprint: https://arxiv.org/pdf/2412.00756

[16] Zhao G, Zhao Y, Yin X, Lin L, Zhu J. Beyond Spurious Cues: Adaptive Multi-Modal Fusion via Mixture-Ofexperts for Robust Sarcasm Detection. Mathematics. 2025;13:3250.

[17] Qin L, Huang S, Chen Q, Cai C, Zhang Y, et al. MMSD2.0: Towards a Reliable Multimodal Sarcasm Detection System. InFindings of the association for computational linguistics: ACL. 2023:10834-10845.

[18] Chen J, Huang S. InterCLIP-MEP: Interactive Clip and Memory-Enhanced Predictor for Multi-Modal Sarcasm Detection. 2024. Arxiv preprint: https://arxiv.org/pdf/2406.16464v1.

[19] Liang B, Lou C, Li X, Yang M, Gui L, et al. Multi-Modal Sarcasm Detection via Cross-Modal Graph Convolutional Network. InProceedings of the 60th annual meeting of the association for computational linguistics. volume 1: long papers. 2022:1767-1777.

[20] Cai Y, Cai H, Wan X. Multi-Modal Sarcasm Detection in Twitter With Hierarchical Fusion Model. InProceedings of the 57th annual meeting of the association for computational linguistics. 2019:2506-2515.

[21] Wang T, Li J, Su G, Zhang Y, Su D, et al. RCLMuFN: Relational Context Learning and Multiplex Fusion Network for Multimodal Sarcasm Detection. 2024. Arxiv preprint: https://arxiv.org/pdf/2412.13008.

[22] Jana S, Kundu A, Singh SR. Think Twice Before You Judge: Mixture of Dual Reasoning Experts for Multimodal Sarcasm Detection. 2025. Arxiv preprint:: https://arxiv.org/pdf/2507.04458

[23] Zhang S, Lian J, Yu G, Xu B, Ao X. Gdcnet: Generative Discrepancy Comparison Network for Multimodal Sarcasm Detection. 2026. Arxiv preprint: https://arxiv.org/pdf/2601.20618

[24] Wang H, Liu F, Jiao L, Wang J, Hao Z, et al. ViLT-CLIP: Video and Language Tuning Clip With Multimodal Prompt Learning and Scenario-Guided Optimization. Proc AAAI Conf Artif Intell. 2024;38:5390-400.

[25] Jana S, Danayak S, Singh SR. Ads: Adapter-State Sharing Framework for Multimodal Sarcasm Detection. 2025. Arxiv preprint: https://arxiv.org/pdf/2507.04508

[26] Kumar A, Vepa J. Gated Mechanism for Attention Based Multi Modal Sentiment Analysis. InICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP. IEEE. 2020:4477-4481.

[27] Gao Y, Liu J, Xu Z, Wu T, Zhang E, et al. Softclip: Softer Cross-Modal Alignment Makes Clip Stronger. Proc AAAI Conf Artif Intell. 2024;38:1860-1868.

[28] Lu X, Ni Y, Ding Z. Cross-Modal Sentiment Analysis Based on Clip Image-Text Attention Interaction. Int J Adv Comput Sci Appl. 2024;15.

[29] He L, Lin Y, Qin G, Liu J, Feng X, et al. Dual Dynamic Multi-Granularity Fusion Method for Multimodal Sarcasm Detection. Neurocomputing. 2026;663:131985.

[30] Khosla P, Teterwak P, Wang C, Sarna A, Tian Y, et al. Supervised Contrastive Learning. 2020. Arxiv preprint: https://arxiv.org/pdf/2004.11362v1