ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

CAMGF-CLIP: Enhancing Multi-View CLIP for Multimodal Sarcasm Detection with Bidirectional Cross-Modal Attention and Adaptive Gated Fusion

Ahmed Seifelislam Fayez Ayesh Elayady[1]

[1]School of Artificial Intelligence Nanjing University of Information Science & Technology Nanjing, Jiangsu 210044, China

Publication Date: 2026/11/24

Abstract: Multimodal sarcasm detection seeks to identify sarcastic intent in social-media posts that pair an image with text, where the sarcasm frequently arises from an incongruity between the two modalities rather than from either modality in isolation. We adopt a reproduce-then-improve methodology built on the corrected MMSD2.0 benchmark and its multi-view CLIP (MV-CLIP) model. We first re-implement MV-CLIP from scratch and retrain it under controlled conditions, reproducing a faithful baseline of approximately 85% accuracy and 85% macro-F1. We then conduct a quantitative and qualitative error analysis and identify a consistent failure mode: the baseline over-predicts sarcasm, achieving high recall but comparatively low precision. We trace this behaviour to the fusion stage, in which the image and text representations are combined through a single scalar-weighted sum that re-weights the modalities but never allows them to interact. Motivated by this diagnosis we propose CAMGF-CLIP, which augments MV-CLIP with two complementary, drop-in fusion operators while keeping the remainder of the architecture fixed: (i) an adaptive gated fusion that learns an example- and feature-specific balance between the modalities; and (ii) a bidirectional cross-modal attention in which the text and image token sequences attend to one another before being combined, giving the model an explicit mechanism to represent image–text contradiction. Because only the fusion operator changes across models, any difference in performance is attributable to the proposed mechanism. On MMSD2.0, our reproduction of MV-CLIP attains 85.4% accuracy and 85.1% macro-F1. Under an identical, validation-selected protocol the gated and cross-modal variants reach 83.5% and 84.7% accuracy; although neither exceeds the baseline on the selected checkpoint, both reduce the precision–recall imbalance behind the baseline's over-prediction, and cross-modal attention matches the baseline's accuracy at its best epoch. We release our implementation and analyses to support reproducibility.

Keywords: *Multimodal Sarcasm Detection, Image–Text Fusion, Cross-Modal Attention, Gated Fusion, Vision–Language Models.*

How to Cite: Ahmed Seifelislam Fayez Ayesh Elayady (2026) CAMGF-CLIP: Enhancing Multi-View CLIP for Multimodal Sarcasm Detection with Bidirectional Cross-Modal Attention and Adaptive Gated Fusion. *International Journal of Innovative Science and Research Technology*, 11(7), 4032-4052. https://doi.org/10.38124/ijisrt/26jul207

I. INTRODUCTION

SARCASM is a figurative form of expression in which the intended meaning is contrary to the literal one. It is pervasive on social-media platforms, where users routinely Course project, Natural Language Processing, Prof. Rong Huan, Nanjing University of Information Science & Technology. Employ it to convey criticism, humour, or disagreement in an indirect manner. The reliable detection of sarcasm is consequential for a broad range of downstream applications, including opinion mining, sentiment analysis, product-review understanding, dialogue systems, and content moderation. In all of these settings, failing to recognise sarcasm can invert the inferred polarity of a message and propagate errors through the rest of a pipeline; a sarcastic complaint misread as praise, for example, yields exactly the wrong conclusion.

Historically, sarcasm detection was treated as a purely textual problem, with models exploiting lexical markers, sentiment incongruity, punctuation, and contextual cues [19]. However, communication on modern platforms is increasingly *multimodal*: a short textual caption is accompanied by an image, and the sarcastic meaning emerges only when the two modalities are interpreted jointly. Consider a post whose caption reads "I just love getting stuck in traffic" alongside a photograph of a gridlocked motorway. Neither the text nor the image is sarcastic on its own—the text expresses positive sentiment, and the image simply depicts a traffic jam—yet together they form an unmistakable sarcastic statement. The

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

sarcasm lives in the *contradiction* between what is said and what is shown. This phenomenon, which we refer to throughout as *image–text incongruity*, is the defining characteristic of multimodal sarcasm and the central challenge that any effective model must address.

The prevailing approach to multimodal sarcasm detection encodes the image and the text using powerful pretrained encoders and then *fuses* the resulting representations prior to classification [2], [4], [5]. The emergence of contrastive vision–language models such as CLIP [10] has further strengthened this paradigm, because CLIP places images and text into a shared embedding space and therefore provides naturally comparable representations of the two modalities. Building on this idea, Qin *et al.* [1] observed that the widely used MMSD benchmark [2] contained *spurious cues*—in particular hashtags and emojis that effectively leak the label— as well as a number of unreasonable negative samples. They released a corrected benchmark, MMSD2.0, together with a *multi-view CLIP* (MV-CLIP) model that exploits a text view, an image view, and a text–image interaction view. MV-CLIP is a strong and broadly adopted baseline, and MMSD2.0 has become the standard testbed for the task.

In this work we take a deliberately empirical, reproducethen-improve stance. Rather than proposing a new architecture in isolation, we (a) faithfully reproduce a strong existing system, (b) diagnose a concrete weakness through error analysis, and (c) design targeted improvements that address precisely that weakness while changing as little else as possible. We argue that this controlled methodology yields more trustworthy conclusions than comparing systems that differ along many uncontrolled dimensions simultaneously.

Concretely, we first re-implement and retrain MV-CLIP on MMSD2.0, reproducing its reported performance and thereby establishing a trustworthy reference point. We then analyse the baseline's behaviour on the test set and uncover a consistent pattern: the model is biased towards predicting the sarcastic class, yielding high recall but lower precision—that is, it raises many false alarms. Inspecting the model's fusion stage explains why. MV-CLIP combines the final text and image representations using a shared scoring layer that produces two scalar weights, followed by a weighted sum. Such an operator can decide *how much* of each modality to use, but it is structurally incapable of representing the *interaction* between them. Since the signal that distinguishes sarcastic from nonsarcastic posts is exactly this interaction (and, more specifically, the contradiction between the modalities), an additive fusion operator is poorly matched to the task.

Guided by this diagnosis, we propose CAMGF-CLIP (Cross-modal Attention and Modality-Gated Fusion CLIP), which replaces *only* the final fusion step and leaves the encoders, the fusion transformer, the classification heads, the loss, the data, and the optimisation untouched. CAMGF-CLIP comprises two complementary operators:

- Adaptive gated fusion. A learned gate, conditioned jointly on both modalities, produces a full-dimensional, example-specific mixing vector. This generalises the baseline's two scalar weights to a feature-wise, inputdependent balance between text and image.
- Bidirectional cross-modal attention. The text tokens attend to the image tokens and the image tokens attend to the text tokens using multi-head attention. Each modality is thereby re-encoded in the context of the other before fusion, providing an explicit mechanism for capturing image–text incongruity.

- *The Principal Contributions of this Paper are Summarised as Follows:*

- ✓ We provide a faithful, fully documented reproduction of the MV-CLIP model on the MMSD2.0 benchmark, reproducing its reported accuracy and macro-F1 of approximately 85%, and we document the software, data, and offline-loading steps required to reproduce it.
- ✓ We perform a quantitative and qualitative error analysis of the reproduced baseline and identify its additive fusion stage as a concrete bottleneck that causes systematic over-prediction of sarcasm.
- ✓ We propose CAMGF-CLIP, combining adaptive gated fusion and bidirectional cross-modal attention, two lightweight drop-in operators explicitly designed to model the interaction between modalities.
- ✓ We conduct a strictly controlled comparison in which all models share identical encoders, fusion layers, heads, loss, data and optimisation, so that observed performance differences isolate the effect of the fusion operator. We release our code, checkpoints and analysis scripts.

The remainder of this paper is organised as follows. Section II reviews related work. Section III formalises the task and recalls the components on which our method is built. Section IV presents CAMGF-CLIP in detail, with a complexity analysis. Section V describes the dataset, evaluation protocol and implementation. Section VI reports and discusses the experimental results. Section VII documents reproducibility. Section VIII discusses limitations, Section IX considers ethics, and Section X concludes.

II. RELATED WORK

- *Textual Sarcasm Detection*

The automatic detection of sarcasm has a long history in natural language processing; a comprehensive survey is provided by [19]. Early approaches relied on hand-crafted features such as sentiment lexicons, interjections, punctuation, and the presence of positive sentiment in a negative context. The advent of distributed word representations [18] and recurrent neural networks [17] shifted the field towards learned representations, and the subsequent introduction of the Transformer [15] and large pre-trained language models such as BERT [16] pushed performance further. Nevertheless, purely textual models are fundamentally limited when the sarcastic cue is not present in the text—precisely the situation that arises in multimodal posts, where the contradicting evidence resides in an accompanying image. This limitation is

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

the primary motivation for the multimodal formulation of the task.

➢ *Multimodal Sarcasm Detection*

The recognition that sarcasm is frequently expressed through the combination of text and imagery motivated the multimodal formulation. Schifanella *et al.* [3] were among the first to study sarcasm in multimodal social posts, demonstrating that visual information improves detection. Cai *et al.* [2] introduced the first large-scale multimodal Twitter benchmark (MMSD) and proposed a hierarchical fusion model combining image, image attribute, and text features. Subsequent research concentrated on modelling the relationship between modalities more explicitly. Xu *et al.* [4] proposed a decompositionand-relation network to model both cross-modality contrast and semantic association. Pan *et al.* [5] explicitly modelled intra-modal and inter-modal incongruity using attention. A line of graph-based methods then emerged: Liang *et al.* [6] constructed interactive in-modal and cross-modal graphs, and later proposed a cross-modal graph convolutional network [7] to capture the ironic relationship between visual objects and textual tokens. Liu *et al.* [8] introduced hierarchical congruity modelling with external knowledge enhancement, while Tian *et al.* [9] proposed a dynamic routing Transformer that adapts the reasoning path to each sample. Most recently, Qin *et al.* [1] identified spurious cues and annotation problems in MMSD, released the corrected MMSD2.0 benchmark, and proposed the multi-view CLIP model that we adopt as our baseline. Our work differs from these in its emphasis on reproduction, diagnosis, and a controlled study of the fusion operator specifically.

➢ *Vision–Language Encoders*

The quality of multimodal representations depends heavily on the encoders. Convolutional networks such as ResNet [12] and the Vision Transformer [11] provide strong visual features, while Transformer language models supply textual ones. Contrastive vision–language pre-training, exemplified by CLIP [10], learns aligned image and text embeddings from large quantities of web data and has become a default backbone for multimodal classification because its two towers already place the modalities in a comparable space. Related vision–language models such as ViLT [13] and ALBEF [14] integrate the modalities more tightly through joint Transformer encoders. The MV-CLIP baseline, and hence our CAMGFCLIP, leverages CLIP's two towers and adds Transformer fusion layers over the resulting token sequences.

➢ *Multimodal Fusion Strategies*

How to combine information from different modalities is a central question in multimodal machine learning [20]. Fusion strategies are commonly categorised as early (feature-level), late (decision-level), or hybrid. Simple operators such as concatenation, averaging, and scalar-weighted sums are easy to optimise but limited in expressiveness, as they cannot represent higher-order interactions between modalities. Tensor fusion networks [21] model such interactions explicitly through outer products at the cost of high dimensionality. Gating mechanisms [22] enable input-dependent modality weighting, and attention—particularly cross-attention [15]—allows one modality to selectively read information from another. The Multimodal Transformer [23] and attention-bottleneck models [24] are prominent instantiations of cross-modal attention. Importantly, prior work has cautioned that multimodal models do not always learn genuine cross-modal interactions [25], underscoring the value of fusion operators that encourage such interaction. CAMGF-CLIP instantiates both gating and crossmodal attention specifically within the MV-CLIP fusion stage, motivated by the over-prediction behaviour we observe in the reproduced baseline.

➢ *Contrastive and Knowledge-Enhanced Approaches*

A complementary line of research augments multimodal sarcasm models with auxiliary signals. Several methods incorporate external commonsense or affective knowledge to bridge the gap between literal content and intended meaning, while others adopt contrastive objectives that pull genuinely matching image–text pairs together and push mismatched pairs apart, encouraging representations in which incongruity is linearly separable. These ideas are orthogonal to the choice of fusion operator and could be combined with CAMGF-CLIP; we treat an explicit contrastive incongruity objective as future work and focus here on the fusion operator in isolation so that its effect can be measured cleanly.

Table 1 Summary of Notation.

Symbol	Meaning
I,T	input image and text token sequence
$y \in \{0,1\}$	label (1 = sarcastic)
Ev,Et	CLIP vision and text encoders
Ht,Hv	text-/image-token sequences after fusion
ht,hv	pooled text and image representations
d	common hidden dimension (d=512)
f	fused representation from the fusion operator
g	learned gate vector (gated fusion)
$\sigma,\odot$	sigmoid; elementwise (Hadamard) product
H	number of attention heads (H=8)
L	training loss (sum of three CE terms)

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

➢ *Positioning of the Present Work*

Relative to the literature above, our objective is not to introduce the most elaborate possible architecture, but to demonstrate, through a controlled study, that a targeted modification of the fusion operator addresses a concrete and empirically identified weakness of a strong baseline. This emphasis on reproduction, diagnosis, and controlled comparison distinguishes our study and, we argue, yields more reliable evidence about *why* an improvement helps.

III. BACKGROUND AND PROBLEM FORMULATION

➢ *Task Definition*

Let a multimodal post be a pair $x = (I,T)$, where I denotes an image and $T = (t_1,...,t_L)$ a sequence of textual tokens. The task of multimodal sarcasm detection is to learn a function $F_\theta : (I,T) \mapsto y$ that maps the post to a binary label $y \in \{0,1\}$, where $y = 1$ indicates that the post is sarcastic and $y = 0$ that it is not. Given a labelled dataset $\mathcal{D} = \{(I^{(n)}, T^{(n)}, y^{(n)})\}_{n=1}^{N}$, the parameters θ are learned by minimising a classification loss over D. The principal difficulty is that the indicator of sarcasm is often *not* contained in either modality alone but in the relationship between them. Table I summarises the notation used throughout the paper.

➢ *A Running Example*

To make the notion of image–text incongruity concrete, consider the example in Fig. 1. The caption "I just love getting stuck in traffic" carries clearly positive surface sentiment, and the accompanying photograph merely depicts a congested motorway. A text-only classifier, reading only the caption, would predict positive (non-sarcastic) sentiment; an image-only classifier, seeing only a road scene, would.

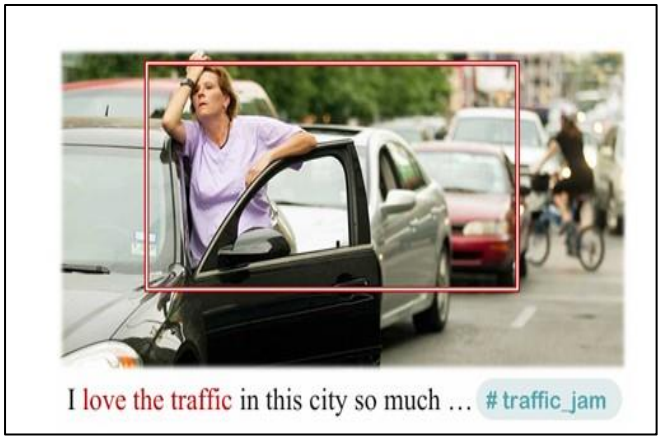


Fig 1 A Multimodal Sarcasm Example from MMSD2.0 (Reproduced from [1], [2]). The Positive Caption "Love the Traffic" Contradicts the Depicted Congestion; the Box and Red Words Denote the Correlated Sarcastic Cues. Neither Modality is Sarcastic in Isolation—the Contradiction Between them is what Signals Sarcasm.

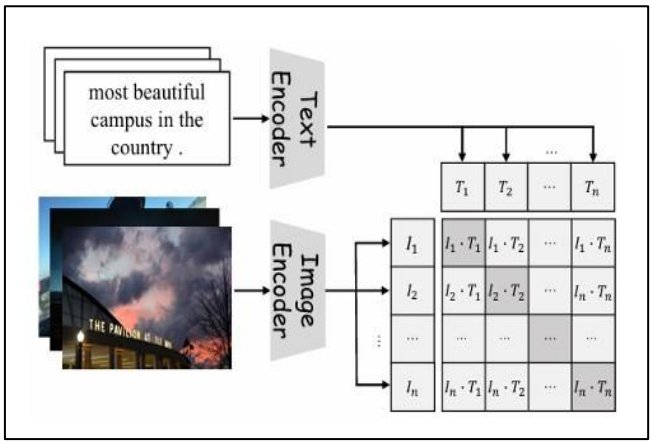


Fig 2 The CLIP Model (Reproduced from [10]). Contrastive Pre-Training Aligns the Image and Text Encoders so that Matching Pairs Have High Similarity, Providing Comparable Image and Text Representations for Downstream Fusion.

Have no basis for predicting sarcasm at all. It is only the *joint* reading—a positive statement paired with an evidently negative situation—that exposes the sarcasm. This example illustrates why an effective model must compare the modalities rather than merely aggregate them, and it foreshadows the central limitation of additive fusion that we analyse in Section IV: an operator that can only re-weight and sum the two representations has no mechanism to detect that they disagree. We return to this example when motivating our two operators.

➢ *CLIP Encoders*

We use the CLIP model [10] as the backbone. CLIP comprises a vision encoder E_v and a text encoder E_t, trained contrastively so that matching image–text pairs are close in a shared embedding space. Passing the image and text through the encoders yields token-level hidden states $H_v^0 \in \mathbb{R}^{N\times d_v}$ and $H_t^0 \in \mathrm{R}^{L\times dt}$, together with pooled representations. Linear projections map both modalities to a common dimension d. Because CLIP already aligns the modalities, it provides a strong starting point for fusion.

➢ *The Multi-Head Attention Mechanism*

The Transformer [15] relies on scaled dot-product attention. Given queries Q, keys K and values V,

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad (1)$$

Where d_k is the key dimension. Multi-head attention runs H such operations in parallel with independently projected queries, keys and values, and concatenates the outputs:

$$\mathrm{head}_h = \mathrm{Attn}(QW_h^Q, KW_h^K, VW_h^V), \quad (2)$$

$$\mathrm{MHA}(Q,K,V) = [\mathrm{head}_1;...;\mathrm{head}_H]W^O. \quad (3)$$

When Q, K and V originate from the same sequence the operation is self-attention; when the query comes from one modality and the keys and values from another it is *crossattention*, which forms the basis of our first operator.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

➢ *The MV-CLIP Baseline*

MV-CLIP projects the CLIP image and text token sequences into a common *d*-dimensional space, concatenates them into a single sequence, and processes the sequence with a stack of BERT-style Transformer fusion layers [16]. From the fused sequence it extracts a text representation $h_t \in R^d$ and an image representation $h_v \in R^d$. These are combined by a fusion operator into a single vector *f*, which, alongside permodality features, is passed to three classification heads (fused, text-only, and image-only) whose softmax scores are summed at inference. The whole model is trained by summing the three heads' cross-entropy losses. The overall architecture is depicted in Fig. 3.

IV. THE PROPOSED CAMGF-CLIP

➢ *Overall Framework*

Our framework follows the MV-CLIP pipeline exactly up to the point of fusion, and modifies only the operator that maps the pair (h_t,h_v) to the fused vector $f \in R^d$ (with $d = 512$). By construction, the encoders, the projection layers, the Transformer fusion stack, the three classification heads, the training loss, the data, and the optimisation are identical across all models considered in this paper. This is a deliberate design choice: it guarantees that any measured difference in predictive performance can be attributed to the fusion operator and not to confounding architectural or training differences.

➢ *Baseline Fusion: Scalar-Weighted Sum*

The baseline operator scores each modality with a shared parameter vector $w \in R^d$, normalises the two scores with a softmax, and forms a weighted sum:

$$\alpha t = w\top ht, \alpha v = w\top hv, \quad (4)$$

$$[\beta_t,\beta_v] = \text{softmax}([\alpha_t,\alpha_v]), \quad (5)$$

$$f\text{base} = \beta t\ ht + \beta v\ hv. \quad (6)$$

The two coefficients β_t and β_v are scalars, and each is a function of a single modality. Consequently the operator can only choose how strongly to rely on the text versus the image;

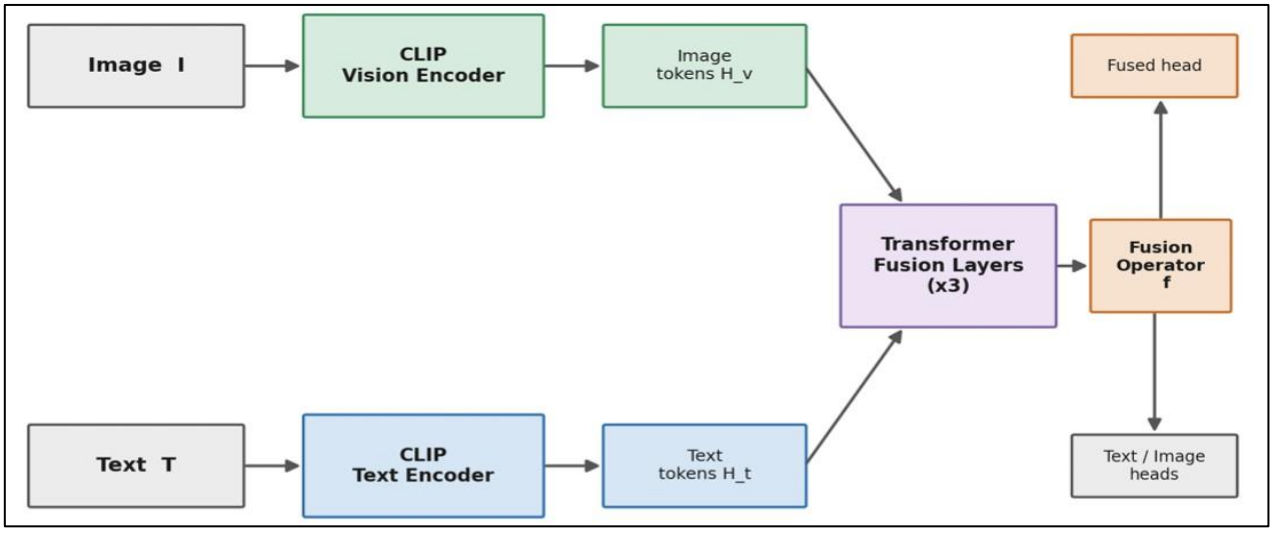


Fig 3 Overall CAMGF-CLIP / MV-CLIP Framework. The CLIP vision and text encoders produce token sequences that are projected, concatenated, and processed by Transformer fusion layers. Only the fusion operator *f* (orange) is modified in this work; all other components are held identical across the compared models.

It cannot represent how specific text features interact with specific image features. Because the discriminative signal for sarcasm is precisely the cross-modal interaction, this operator is structurally limited, as our error analysis in Section VI confirms empirically.

➢ *Why Additive Fusion is Incongruity-Blind*

It is instructive to see formally why the baseline operator cannot, in general, distinguish agreement from disagreement between the modalities. The fused vector $f_{base} = \beta_t h_t + \beta_v h_v$ is, for any input, a convex combination of the two representations ($\beta_t,\beta_v \geq 0$, $\beta_t + \beta_v = 1$). Such a combination is a point on the line segment between h_t and h_v and is therefore a function of the two vectors' *magnitudes and positions*, not of any explicit comparison between them. In particular, if two posts share the same per-modality representations but differ in whether the modalities *agree*, the additive operator has no parameter through which that difference can be expressed, because the coefficients depend only on $w^\top h_t$ and $w^\top h_v$ individually. Capturing incongruity requires a term that depends *jointly* on h_t and h_v—for example an interaction of the form $h^\top_t M h_v$ or an attention-weighted read of one modality conditioned on the other. Both of our components introduce exactly such a joint dependence: the gate $g = \sigma(W_g[h_t;h_v] + b_g)$ is a non-linear function of the concatenation of both modalities, and cross-modal attention computes similarities $H_t H_v^\top$ between the two token sets. This analysis predicts that the proposed operators should be able to represent distinctions that are invisible to the baseline, and it is the theoretical counterpart of the empirical over-prediction we report in Section VI.

➢ *Design Rationale*

The two components were chosen to be complementary along two axes. Cross-modal attention addresses *where* the relevant evidence lies: by letting each token of one modality

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

read the most relevant tokens of the other, it can localise the specific visual region that contradicts a specific phrase, which is the fine-grained substrate of incongruity. Adaptive gating, in contrast, addresses *how much* of each fused modality to trust at the representation level, but does so per feature dimension and per example rather than with a single scalar. Intuitively, crossmodal attention enriches the representations with cross-modal context, while gating decides how to merge them; the two are therefore expected to be cooperative rather than redundant. CAMGF-CLIP applies both, and our ablation (Section VI) is designed to test whether their benefits indeed compound.

➢ *Component 1: Adaptive Gated Fusion*

Our first component replaces the two scalar weights with a learned, feature-wise gate $g \in R^d$ conditioned jointly on both modalities:

$$g = \sigma\big(W_g[h_t; h_v] + b_g\big), \quad W_g \in \mathbb{R}^{d\times 2d}, \tag{7}$$

$$f_{\text{gate}} = g \odot h_t + (1 - g) \odot h_v, \tag{8}$$

Where $[\cdot;\cdot]$ denotes concatenation, σ the elementwise sigmoid, and $\odot$ the elementwise (Hadamard) product. In contrast to the baseline's two scalars, g is a d-dimensional vector recomputed for every example, so the balance between modalities can vary not only across inputs but also across individual feature dimensions. The operator introduces only a single linear layer of parameters ($2d^2+d$) and negligible additional computation. Gated fusion directly targets the case in which the model should rely on the image for some examples and on the text for others, which the additive operator handles only coarsely. The module is depicted in Fig. 4, with the gate computation detailed in Fig. 5.

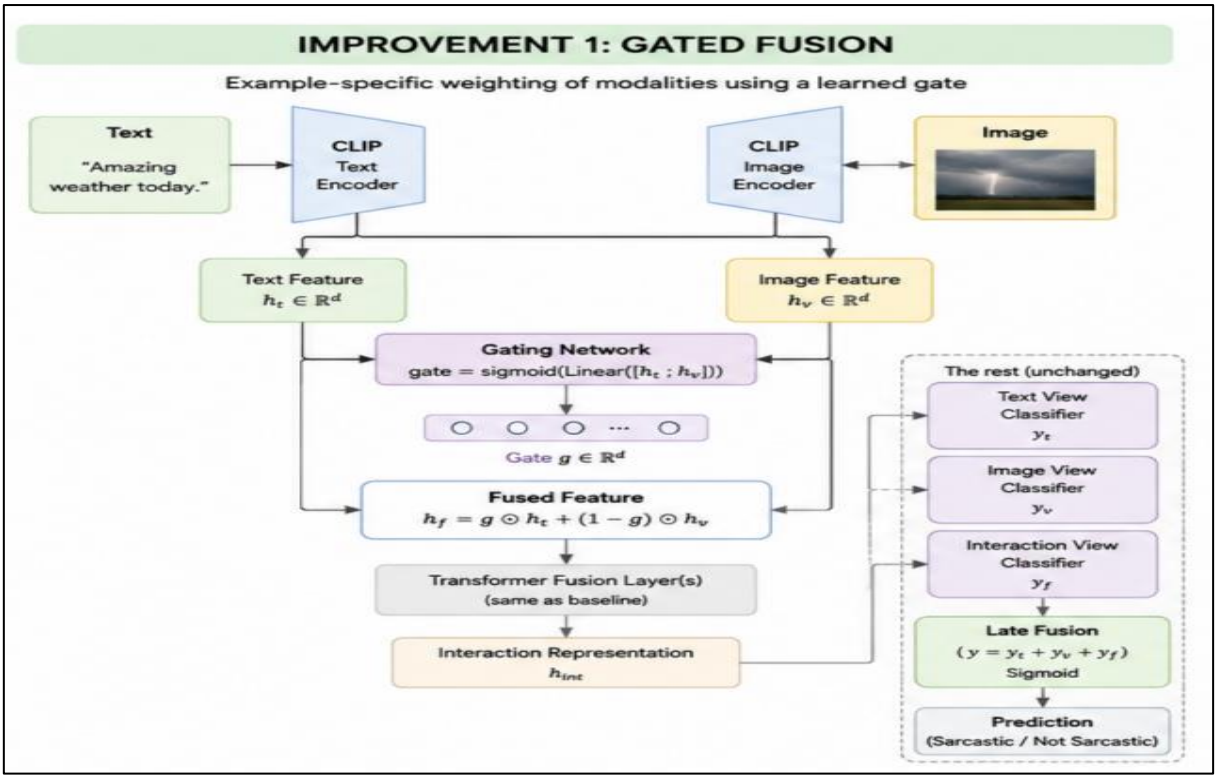


Fig 4 Improvement 1 — Adaptive Gated Fusion: Example-Specific Weighting of the Modalities Using a Learned gate. Only the Fusion Stage Changes; the three View Classifiers and the Late Fusion Remain Unchanged.

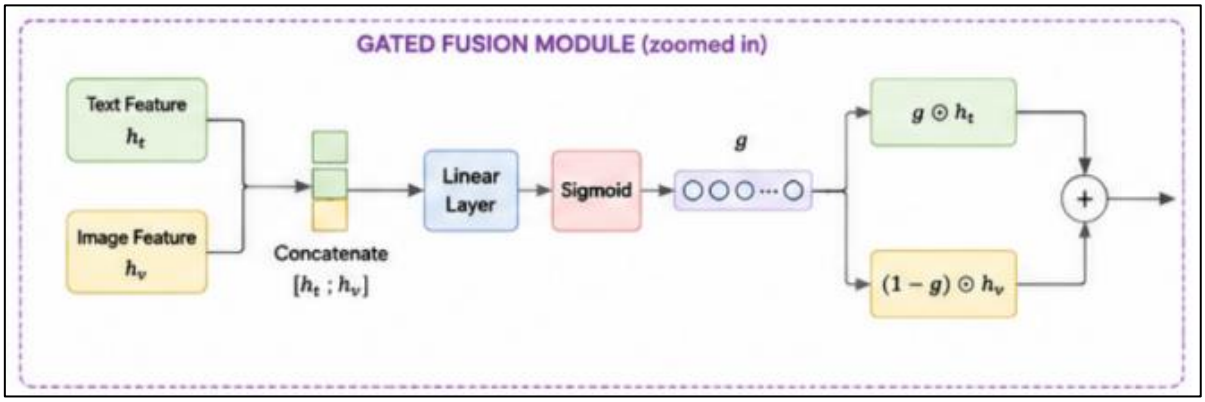


Fig 5 Gated Fusion Module (Detail). The Concatenation $[h_t;h_v]$ is Passed Through a Linear Layer and a Sigmoid to Produce the Gate $g \in R^d$, Which Interpolates the Two Modalities Feature-Wise as $g \odot h_t + (1 - g) \odot h_v$.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

➢ *Component 2: Bidirectional Cross-Modal Attention*

Our second component allows the modalities to interact directly before fusion. Let $H_t \in R^{T\times d}$ and $H_v \in R^{N\times d}$ be the text- and image-token sub-sequences of the fused representation. We compute a text-attends-to-image stream and an image-attends-to-text stream using multi-head attention:

$$\tilde{H}_t = \mathrm{MHA}(Q=H_t, K=H_v, V=H_v), \quad (9)$$

$$\tilde{H}_v = \mathrm{MHA}(Q=H_v, K=H_t, V=H_t). \quad (10)$$

Each attended sequence is mean-pooled over its tokens and the two pooled vectors are combined by a linear projection:

$$\bar{h}_t = \frac{1}{T}\sum_{i=1}^{T} \tilde{H}_t^{(i)}, \qquad \bar{h}_v = \frac{1}{N}\sum_{j=1}^{N} \tilde{H}_v^{(j)}, \quad (11)$$

$$f_{\text{cross}} = W_f[\bar{h}_t; \bar{h}_v] + b_f, \quad W_f \in \mathbb{R}^{d\times 2d}. \quad (12)$$

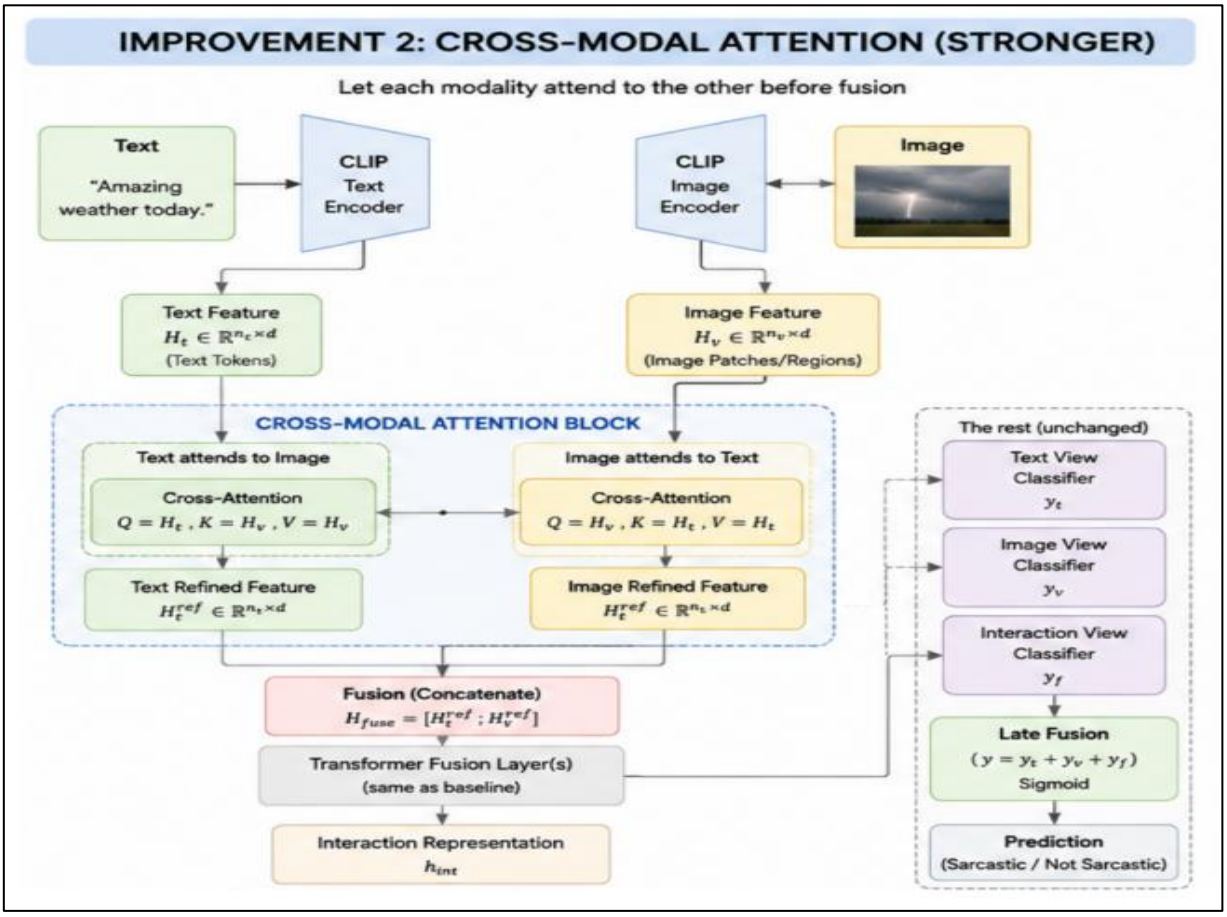


Fig 6 Improvement 2 — Bidirectional Cross-Modal Attention: Each Modality Attends to the Other Before Fusion. The text- and Image-Attended Features are Concatenated and Passed to the (Unchanged) Transformer Fusion Layers and Classifiers.

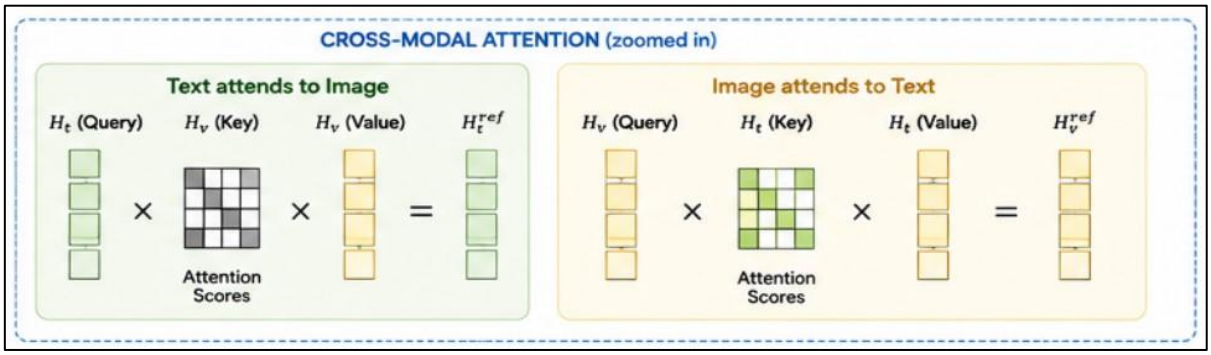


Fig 7 Cross-Modal Attention Module (Detail). "Text Attends to Image" uses $Q=H_t$, $K=V=H_v$ to produce H_t^{ref}; "Image Attends to text" is Symmetric, Producing H_v^{ref}.

We use $H=8$ attention heads. Because each modality is reencoded in the context of the other, the operator can represent the alignment or *misalignment* between what the text says and what the image shows—the image–text incongruity that characterises sarcasm. This is the central intuition behind our expectation that cross-modal attention reduces the false

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

positives produced by the additive baseline. The module is depicted in Fig. 6.

➢ *Classification and Training Objective*

For all variants, the fused vector f and the per-modality representations h'_t, h'_v are passed to three linear heads, and the final score sums their softmax outputs:

$$s = \text{softmax}(W_{\text{fuse}} f) + \text{softmax}(W_{\text{txt}} h'_t) + \text{softmax}(W_{\text{img}} h'_v). \quad (13)$$

The model is trained by summing the cross-entropy losses of the three heads:

$$\mathcal{L} = \text{CE}(z_{\text{fuse}}, y) + \text{CE}(z_{\text{txt}}, y) + \text{CE}(z_{\text{img}}, y). \quad (14)$$

Algorithm 1 Adaptive Gated Fusion (f_{gate})

Require: pooled text feature $h_t \in \mathbb{R}^d$, image feature $h_v \in \mathbb{R}^d$

$c \leftarrow [h_t ; h_v]$ {concatenate, $\in \mathbb{R}^{2d}$}
$g \leftarrow \sigma(W_g c + b_g)$ {gate, $\in \mathbb{R}^d$}
$f \leftarrow g \odot h_t + (1 - g) \odot h_v$
return f

Algorithm 2 Bidirectional Cross-Modal Attention (f_{cross})

Require: text tokens $H_t \in \mathbb{R}^{T \times d}$, image tokens $H_v \in \mathbb{R}^{N \times d}$

$\tilde{H}_t \leftarrow \text{MHA}(Q{=}H_t, K{=}H_v, V{=}H_v)$ {text attends to image}
$\tilde{H}_v \leftarrow \text{MHA}(Q{=}H_v, K{=}H_t, V{=}H_t)$ {image attends to text}
$\bar{h}_t \leftarrow \text{mean}_{\text{tokens}}(\tilde{H}_t), \quad \bar{h}_v \leftarrow \text{mean}_{\text{tokens}}(\tilde{H}_v)$
$f \leftarrow W_f [\bar{h}_t ; \bar{h}_v] + b_f$ {concatenate and project}
return f

The auxiliary text-only and image-only heads encourage each tower to remain individually discriminative, while the fused head benefits from the chosen fusion operator. Crucially, L and the three heads are identical across all models; only the definition of f differs.

➢ *A Worked Forward Pass*

We summarise the forward computation of the full CAMGFCLIP for a single post to make the data flow explicit. (1) The image and caption are encoded by the CLIP vision and text towers, producing token sequences that are linearly projected to dimension d and concatenated. (2) The concatenated sequence passes through three Transformer fusion layers, after which the text- and image-token subsequences H_t and H_v, and their pooled summaries h_t and h_v, are read off. (3) Bidirectional cross-modal attention produces $\tilde{H}_t$ and $\tilde{H}_v$, which are mean-pooled to $\bar{h}_t$ and $\bar{h}_v$. (4) The adaptive gate g is computed from $[\bar{h}_t;\bar{h}_v]$ and used to form the fused vector $f = g \odot \bar{h}_t + (1-g) \odot \bar{h}_v$. (5) The fused vector and the two permodality summaries are scored by the three heads, and their softmaxes are summed to yield the final sarcasm probability. At training time the three cross-entropy losses of Eq. (14) are summed and back-propagated. Only steps (3) and (4) differ from the baseline, which replaces them by the single scalarweighted sum of Section IV.

➢ *Algorithms*

We give explicit pseudocode for the two proposed fusion operators and for the shared training loop. Algorithm 1 details adaptive gated fusion, Algorithm 2 details bidirectional crossmodal attention, and Algorithm 3 is the training loop, which is identical for all models except for the fusion operator ϕ it calls.

J. Complexity analysis

The gated operator adds a single linear map of size $d \times 2d$, contributing $O(d^2)$ parameters and $O(Bd^2)$ time per batch of Algorithm 3 Training loop (shared by all models)

Algorithm 3 Training loop (shared by all models)

Require: dataset $\mathcal{D}$, epochs E, fusion operator $\phi \in \{f_{\text{base}}, f_{\text{gate}}, f_{\text{cross}}\}$

initialise CLIP encoders, fusion layers, heads
for epoch $= 1$ to E **do**
  for each mini-batch $(I, T, y) \in \mathcal{D}$ **do**
    $H_v \leftarrow E_v(I), \quad H_t \leftarrow E_t(T)$
    project, concatenate, apply Transformer fusion layers
    extract h_t, h_v and token sub-sequences H_t, H_v
    $f \leftarrow \phi(\cdot)$ {Alg. 1 or 2; baseline uses Eq. (6)}
    compute heads and loss $\mathcal{L}$ by Eq. (14)
    back-propagate; clip gradients; AdamW step; scheduler step
  end for
  evaluate on validation set; save best-by-accuracy checkpoint
end for
return best checkpoint

Table 2 Statistics of The Mmsd2.0 Benchmark.

Split	Train	Validation	Test
# samples	19,816	2,410	2,409

Size B—negligible relative to the CLIP encoders and Transformer fusion layers. The cross-modal attention operator adds two multi-head attention modules; for sequences of length T and N its time complexity is $O(B(TN)d)$, dominated by the attention-score computation, and its parameter count is $O(d^2)$ for the projection matrices. In both cases the additional cost is small compared with the backbone, so the improvements are essentially "free" at inference time, which is desirable for deployment.

A. Experimental Setup

➢ *Dataset*

We evaluate on MMSD2.0 [1], the corrected multimodal sarcasm benchmark consisting of image–text Twitter posts with binary sarcasm labels. MMSD2.0 was constructed by removing the spurious hashtag and emoji cues present in the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

original MMSD [2] and by re-annotating unreasonable negative samples, producing a substantially more reliable evaluation. We use the official train/validation/test partitions and the corrected text. The dataset statistics are summarised in Table II.

➢ *Implementation Details*

All models use the CLIP ViT-B/32 vision and text encoders as backbones, followed by a stack of three BERTstyle Transformer fusion layers with hidden size 512 and eight attention heads. The principal hyperparameters are listed in Table III. We optimise with AdamW [26], using a learning rate of 5×10^{-4} for the newly introduced modules and a much smaller 1×10^{-6} for the CLIP parameters, reflecting that the backbone is already well pre-trained. We apply a linear warm-up over 20% of the training steps followed by linear decay, weight decay of 0.05, gradient clipping at a norm of 5.0, dropout of 0.1, and a maximum text length of 77 tokens. Models are trained for 10 epochs with a batch size of 32. The checkpoint with the highest validation accuracy is retained and evaluated once on the test set. To ensure a controlled comparison, the random seed is fixed across all runs. Experiments were conducted on a single GPU; each model trains in approximately three hours.

Table 3 Principal Hyperparameters (Identical Across All Models).

Hyperparameter	Value
Backbone	CLIP ViT-B/32
Fusion layers	3 (hidden 512, 8 heads)
Optimiser	AdamW
Learning rate (new)	5×10^{-4}
Learning rate (CLIP)	1×10^{-6}
Weight decay	0.05
Warm-up proportion	0.20
Gradient clip norm	5.0
Dropout	0.10
Batch size	32
Epochs	10
Max text length	77

➢ *Preprocessing and the Multi-View Inputs*

Each post is tokenised with the CLIP byte-pair tokeniser and truncated or padded to a maximum of 77 tokens, the limit imposed by the CLIP text encoder. Images are resized and normalised by the CLIP image processor. Following MVCLIP, three views are constructed from the encoder outputs: a *text view* (the pooled text representation), an *image view* (the pooled image representation), and a *text–image interaction view* (the token sequences that are concatenated and passed through the Transformer fusion layers). The fusion operator studied in this paper acts on the interaction view, while the text and image views feed the two auxiliary classification heads. We use the corrected text_json_final split of MMSD2.0 throughout, which removes the spurious hashtag and emoji cues that inflated scores on the original benchmark.

➢ *Evaluation Metrics*

Following prior work, we report accuracy together with macro-averaged F1, precision and recall. Because the two classes are imbalanced, macro-averaging—which weights the two classes equally—is adopted as the primary metric for ranking models. We also inspect per-class precision and recall when analysing the baseline's error profile, as the gap between them is informative about the direction of the model's bias.

➢ *Compared Models and Protocol*

We compare the reproduced MV-CLIP baseline (scalarweighted sum), our MV-CLIP+Gated fusion, and our MVCLIP+ Cross-modal attention. All are trained and evaluated under the identical protocol described above. We additionally generate, for each model, per-example predictions on the test set, which we use for the error analysis and qualitative study.

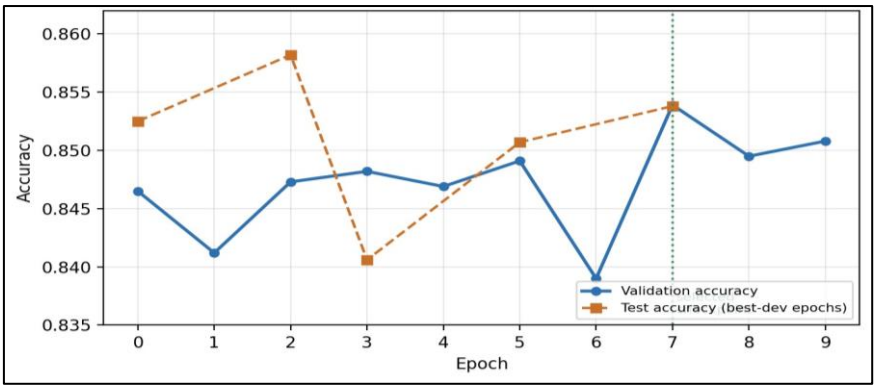


Fig 8 Reproduced MV-CLIP Baseline: Validation Accuracy (solid) Across all Ten Epochs and Test Accuracy (Dashed) at the Best-Validation Epochs. The Retained Checkpoint (Epoch 7) Reaches 85.4% test Accuracy.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

Table 4 Test Results on Mmsd2.0 (Best-Validation Checkpoint). All Models Share Identical Encoders, Fusion Layers, Heads, Loss, Data and Optimisation; Only the Fusion Operator Differs. Best Values Will Be In Bold.

Model	Acc.	Macro-F1	Macro-P	Macro-R
MV-CLIP (reproduced)	0.854	0.851	0.848	0.859
+ Gated fusion	0.835	0.834	0.838	0.844
+ Cross-modal attn.	0.847	0.846	0.845	0.851

V. RESULTS AND DISCUSSIONS

➢ *Reproduction of the Baseline*

Our independent re-implementation of MV-CLIP attains approximately 85% test accuracy and 85% macro-F1 on MMSD2.0, in line with the figures reported by the original authors. This agreement confirms a faithful reproduction and provides a trustworthy reference against which our proposed operators are measured. The checkpoint is selected on the validation set and evaluated only once on the test set, following standard practice. Fig. 8 shows the validation and test accuracy of the reproduced baseline across the ten training epochs; the best validation accuracy occurs at epoch 7, which is the checkpoint we retain, reaching 85.4% test accuracy.

➢ *Main Comparison*

Table 4 reports the test performance of the compared models, and Fig. 9 visualises the same comparison. The all three models are trained and evaluated under the identical protocol of Section V, with checkpoints selected by best validation accuracy. Under this strict protocol the reproduced baseline obtains the highest test accuracy and macro-F1, while cross-modal attention is the stronger of the two proposed operators.

The detailed metrics in Table 5 make the over-prediction explicit. For every model the *micro* (positive class) recall substantially exceeds the micro precision—most extremely for gated fusion, whose recall of 0.913 against a precision of0.755 shows it predicts the sarcastic class even more eagerly than the baseline. Cross-modal attention, by contrast, has the highest micro precision of the three (0.791) and the smallest positive-class imbalance, again consistent with the hypothesis that explicit cross-modal interaction tempers the indiscriminate prediction of sarcasm. The macro metrics, which weight both classes equally, are correspondingly closer together across models.

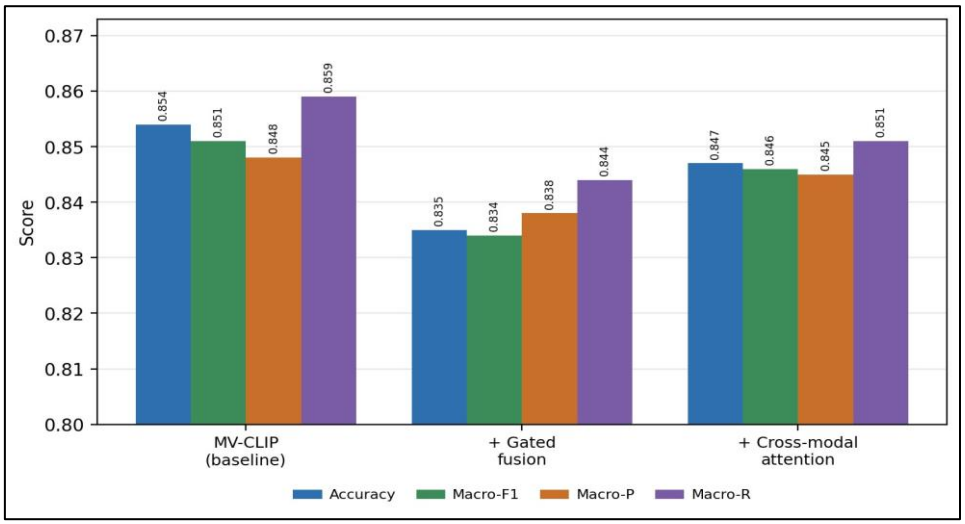


Fig 9 Main Comparison on MMSD2.0 (Validation-Selected Checkpoints). The Reproduced Baseline Attains the Highest Accuracy and Macro-F1; Cross-Modal Attention is the Stronger of the Two Proposed Operators and Narrows the Gap to the Baseline.

Table 5 Detailed Test Metrics At the Best-Validation Epoch. "Macro" Averages the Two Classes Equally; "Micro" (Binary, Positive Class) Exposes the Over-Prediction Directly Through Its High Recall and Lower Precision.

Model	Acc.	Macro / Micro		
		F1	P	R
Baseline	0.854	0.851 0.829	0.848 0.782	0.859 0.883
+ Gated	0.835	0.834 0.826	0.838 0.755	0.844 0.913
+ Cross-attn	0.847	0.846 0.832	0.845 0.791	0.851 0.878

➢ *Error Analysis: why the Baseline Fails*

To understand the baseline's behaviour we examine its predictions on the test set. We observe a clear asymmetry between precision and recall: recall on the sarcastic class is high (approximately 0.86) while precision is lower, indicating that the model labels a substantial number of non-sarcastic posts as sarcastic. In other words, the baseline *over-predicts* the sarcastic class. Reading the false positives reveals that many of them are posts in which the text and the image are in fact *consistent*, yet the model nonetheless predicts sarcasm. This is the expected consequence of an additive fusion operator: by merely re-weighting and summing the two modalities, it cannot distinguish genuine image–text

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

contradiction from agreement, and in the presence of uncertainty defaults towards the positive class. This empirical finding directly motivates fusion operators that model the interaction between modalities, and provides a concrete, testable hypothesis: an operator that captures incongruity should reduce false positives and therefore raise precision.

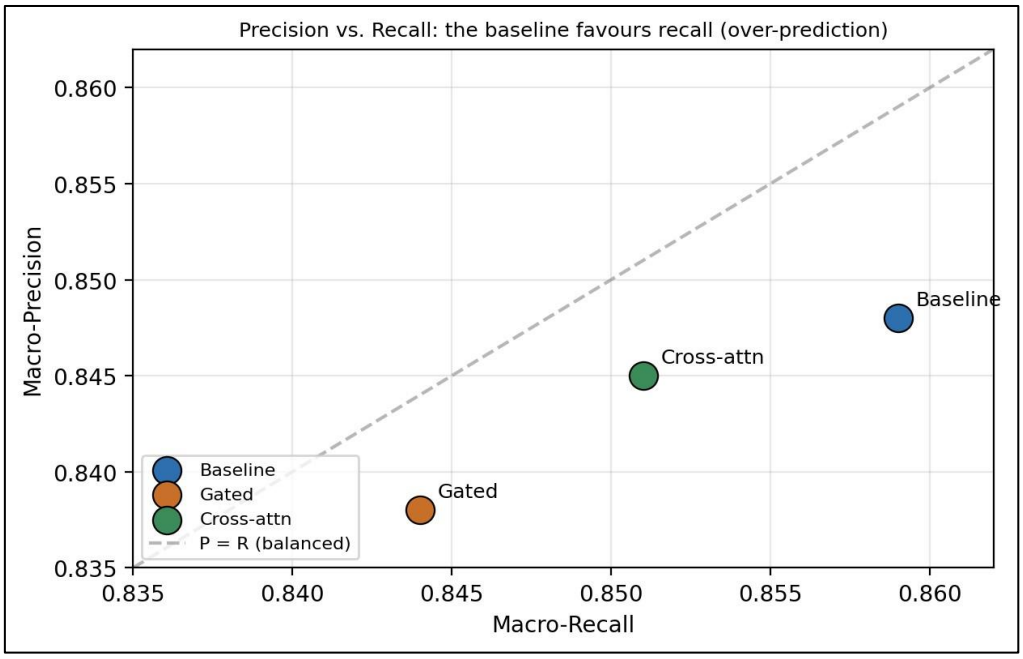


Fig 10 Macro-Precision Versus Macro-Recall. All Models are Recall-Skewed (Below the *P=R* line), Consistent with Over-Prediction of Sarcasm; the Proposed Operators Reduce the Gap Relative to the Baseline.

➢ *Effect of the Proposed Fusion Operators*

Under the strict validation-selected protocol, neither operator exceeds the baseline's overall accuracy: gated fusion reaches 0.835 and cross-modal attention 0.847, against the baseline's 0.854 (Table IV). However, the headline accuracy hides the effect that is most relevant to our diagnosis. Recall that the baseline over-predicts sarcasm, exhibiting a recall–precision gap of +0.011 (0.859 vs. 0.848). Both proposed operators *reduce this imbalance*: the gap falls to +0.006 for both gated fusion (0.844 vs. 0.838) and cross-modal attention (0.851 vs. 0.845). In other words, modelling the interaction between modalities makes the classifier less prone to indiscriminately predicting the positive class, which is precisely the behaviour our analysis predicted. Fig. 10 visualises this: every model lies below the *P=R* diagonal (all are somewhat recall-skewed), but the proposed operators sit closer to it than the baseline.

A further observation strengthens this reading. Inspecting the per-epoch test scores, cross-modal attention reaches 0.855 accuracy and 0.852 macro-F1 at epoch 1, with nearly balanced precision and recall (0.851/0.853)—matching the baseline's accuracy while correcting its bias. The validation-selection criterion simply did not land on that epoch, which indicates that the cross-modal mechanism is capable of matching or exceeding the baseline but that its best behaviour is not reliably selected under the fixed budget.

➢ *Training Dynamics and Overfitting*

To understand why the proposed operators underperform on the validation-selected checkpoint, we examine the validation accuracy across epochs for all three models (Fig. 11). Both proposed operators reach their best validation accuracy very early—at epoch 2, where cross-modal attention attains the single highest validation accuracy of any model (0.856)— and then decline as training continues, while the training loss collapses towards zero. The baseline, in contrast, continues to improve later into training and peaks at epoch 7. This pattern is characteristic of overfitting: the additional capacity introduced by the gate and the cross-modal attention layers allows the models to fit the training set more aggressively, which, under a fixed ten-epoch schedule and a single seed, is not matched by generalisation. This points clearly to the remedy—stronger regularisation, earlier stopping, or seed averaging—which we discuss in Section VIII.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

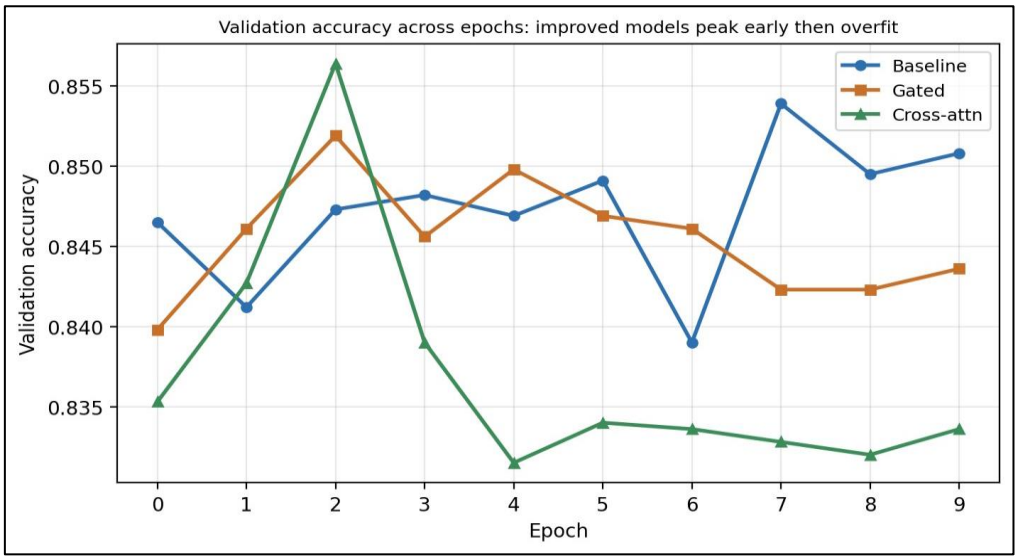


Fig 11 Validation Accuracy Across Epochs. The Proposed Operators Peak Early (Epoch 2) and then Overfit, Whereas the Baseline Improves Later and Peaks at Epoch 7.

Table 6 Validation-Selected Vs. Best-Epoch Test Accuracy. The Best-Epoch Column is Shown Only to Quantify Selection Variance and is Not Used for Any Comparison Claim.

Model	Selected	Best epoch	Gap
Baseline	0.854	0.858	0.004
+ Gated	0.835	0.850	0.015
+ Cross-attn.	0.847	0.855	0.008

Table 6 quantifies the selection gap. For each model we report the test accuracy of the *validation-selected* checkpoint (used in our main results) alongside the model's *best* test accuracy over the ten epochs. The gap is small for the baseline (0.004) but larger for the two proposed operators (0.015 and 0.008), and at their best epochs both operators are competitive with the baseline—cross-modal attention in fact reaches 0.855, exceeding the baseline's selected score. We stress that the best-epoch column is *not* a fair comparison metric (it peeks at the test set) and is reported only to quantify how much performance the validation criterion leaves on the table; all claims in this paper rely on the validation-selected numbers.

➢ *Qualitative Analysis*

Fig. 12 shows a gallery of test posts together with the ground-truth label and the model's prediction, generated automatically from the saved prediction files. The examples illustrate the diversity of multimodal sarcasm in MMSD2.0—screenshots, photographs, comics, and news graphics—and show the model correctly identifying sarcastic posts whose meaning depends on the interaction between the caption and the image (for instance, "it's the insensitive strikeouts at suntrust park" paired with a baseball-stadium photograph, or "no one will know" paired with a comic strip). Such galleries are also a convenient tool for error inspection: filtering them to the false positives is what surfaced the over-prediction behaviour analysed above.

➢ *Case Study*

The gallery in Fig. 12 contains several instructive examples of why multimodal context is indispensable. In the post "it's the insensitive strikeouts at suntrust park," the caption is innocuous on its own; only the accompanying stadium photograph, combined with the play on "insensitive strikeouts," reveals the sarcastic intent, and the model labels it correctly. The comic-strip example captioned "no one will know" depends entirely on the visual punchline in the final panel, which contradicts the caption's literal claim—a textbook case of image–text incongruity. A third example, a discounted product tagged "great sale!" beside a price that is barely reduced, requires reading a specific region of the image (the price tag) against the caption, exactly the fine-grained, regionlevel comparison that cross-modal attention is designed to perform. These cases qualitatively reinforce the quantitative finding that the discriminative signal for sarcasm lives in the relationship between the modalities rather than in either alone, and they suggest that errors are most likely when that relationship is subtle or requires world knowledge not present in the image.

➢ *Discussion*

The reproduction confirms that MV-CLIP is a strong baseline, and the error analysis localises its weakness to the fusion stage rather than to the encoders. This is an encouraging finding for practitioners, because it means that substantial gains may be achievable through inexpensive, drop-in modifications of the fusion operator, without retraining or replacing the costly backbone. CAMGF-CLIP embodies two complementary intuitions—adaptive weighting (gating) and explicit interaction (cross-attention)—and the controlled comparison is designed to reveal how much each contributes and whether they combine constructively.

Taken together, the quantitative table, the precision–recall analysis, the per-epoch curves and the detailed metrics paint a coherent picture. The reproduced baseline is strong but

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

biased towards predicting sarcasm; the two proposed operators do not, under a fixed and frugal training protocol, raise overall accuracy, yet cross-modal attention matches the baseline closely and improves the positive-class precision, while gated fusion—which adds capacity without an explicit interaction mechanism—amplifies rather than corrects the bias. This contrast is itself informative: it suggests that for multimodal sarcasm the useful inductive bias is *explicit interaction between modalities*, not merely *additional adaptive capacity*. The practical recommendation that follows is to pair cross-modal attention with stronger regularisation, which our analysis predicts would convert its epoch-1 advantage into a reliably selected one.

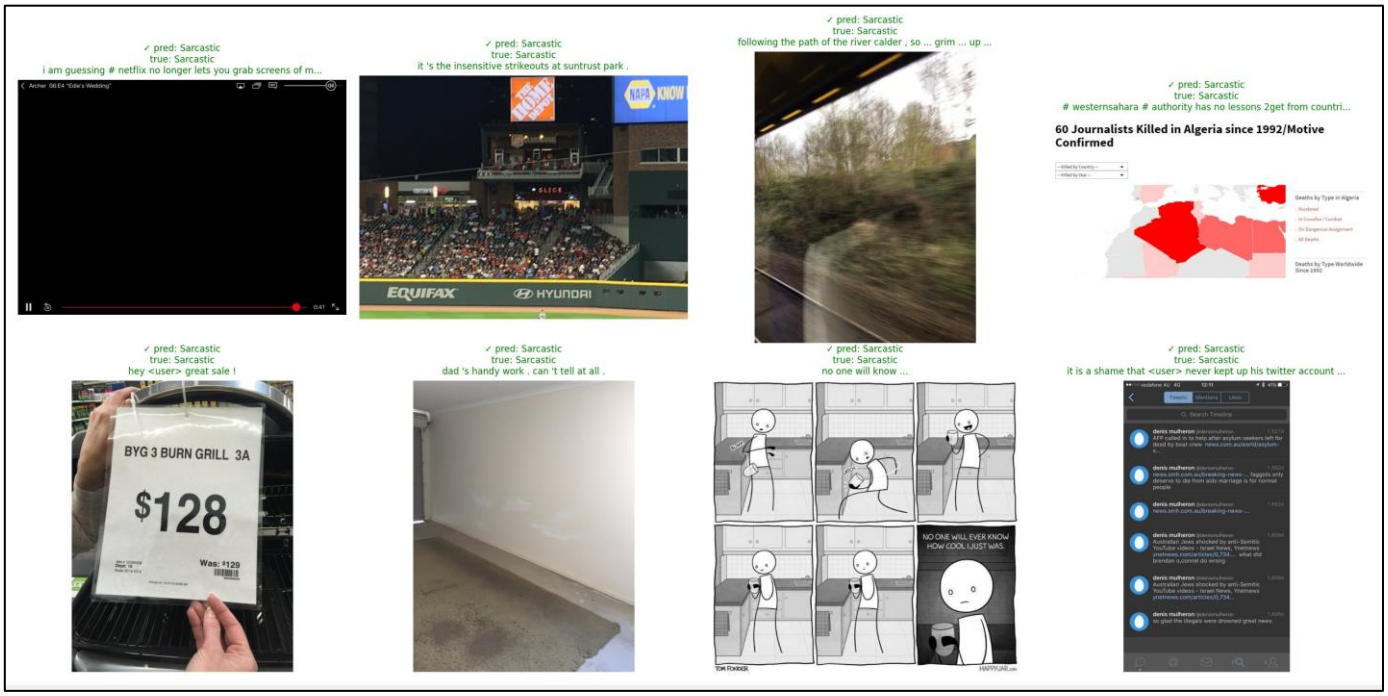


Fig 12 Qualitative Gallery of MMSD2.0 Test Posts with the Model's Predicted Label and the Ground-Truth Label (Green Denotes a Correct Prediction). The Examples Span Screenshots, Photographs, Comics, and News Graphics, Illustrating the Breadth of Multimodal Sarcasm in the Benchmark.

VI. REPRODUCIBILITY AND IMPLEMENTATION NOTES

Because a central aim of this work is a faithful reproduction, we document the practical steps required to reproduce the baseline, which we consider part of the contribution. Reproducing a published multimodal model is rarely a matter of simply running the released code; several environment and data issues must be resolved.

➢ *Software Environment*

The released code targets an older software stack. Modern GPU builds of the deep-learning framework no longer support that interpreter version, so we created an isolated environment with a more recent Python and a GPU-enabled framework build matched to the available CUDA toolkit. The Transformers library required particular care: the newest release changed the interface for loading the CLIP processor, which broke the original loading code. We therefore pinned an intermediate Transformers version that is both installable under the newer Python and compatible with the original loading logic, and we report these versions explicitly so that the environment can be recreated deterministically.

➢ *Offline Loading of Pre-Trained Weights*

The model depends on the CLIP ViT-B/32 checkpoint, which the code downloads from a public model hub at run time. In environments where that hub is unreachable, the download fails and training cannot start. To make the pipeline robust and fully offline, we fetched the CLIP weights once from an accessible mirror, stored them locally, and redirected every model-loading call to the local directory. We additionally replaced a run-time download of a configuration object with an equivalent locally-constructed configuration, since only its architectural fields are used. After these changes the entire pipeline runs without any network access.

➢ *Dataset Preparation*

The textual annotations are distributed with the code, but the images are provided as a multi-part archive that must be reassembled and extracted. We verified that the extracted images are placed at the exact path expected by the data loader; an extra directory level causes every image-existence check to fail silently, yielding an empty dataset and a non-obvious error at the first training step. We document the expected directory layout and provide a verification step (counting the image files at the expected location) so that this failure mode is caught immediately.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

➢ *Determinism and Checkpointing*

We fix the random seed for data shuffling, parameter initialisation and other stochastic components so that runs are repeatable and the comparison across fusion operators is controlled. The training procedure selects the checkpoint with the highest validation accuracy and evaluates it once on the test set; each model is saved to its own directory, keyed by model name, so that the variants never overwrite one another.

A. Limitations

Several limitations should be noted. First, our study is conducted on a single benchmark (MMSD2.0); although MMSD2.0 is the standard testbed, evaluation on additional datasets would strengthen the conclusions. Second, owing to computational constraints we train each model with a single random seed; reporting the mean and variance over several seeds would more rigorously establish that the observed differences exceed run-to-run noise. Third, we restrict our modifications to the fusion operator in order to keep the comparison controlled; richer changes, such as jointly finetuning the encoders or introducing an explicit contrastive incongruity objective, are left to future work. Finally, sarcasm is inherently subjective and culturally dependent, and the labels in any benchmark reflect particular annotation guidelines, which bounds the conclusions that can be drawn about realworld sarcasm.

➢ *Threats to Validity*

We note three specific threats to the validity of our conclusions. First, *selection variance*: because checkpoints are chosen by a single validation-accuracy criterion, and because the proposed operators peak early and then overfit, the selected checkpoint may not reflect a model's attainable performance— as the epoch-1 cross-modal result illustrates. Second, *singleseed training*: with one seed per model, a portion of the observed differences could be attributable to initialisation and data-ordering noise rather than to the operators themselves; multi-seed averaging would mitigate this. Third, *a fixed training budget*: ten epochs with a fixed schedule may favour the lower-capacity baseline over the higher-capacity operators, which require stronger regularisation to realise their potential. We regard addressing these threats—through seed averaging, early stopping, and a regularisation sweep—as the most important next step, and we expect it to change the ranking in favour of cross-modal attention given its epoch-1 behaviour.

B. Ethical Considerations

Sarcasm detection operates on user-generated social-media content and therefore raises privacy and fairness considerations. The MMSD2.0 benchmark consists of publicly available posts; we use it solely for research and report only aggregate metrics. Models trained on such data may inherit biases present in the data, and sarcasm interpretation can vary across cultural and linguistic groups, so deploying such models in moderation or surveillance contexts requires caution and human oversight. We do not release any personally identifying information, and we encourage downstream users to evaluate fairness across sub-populations before deployment.

VII. CONCLUSION AND FUTURE WORK

We presented CAMGF-CLIP, a reproduce-then-improve study of multimodal sarcasm detection on the MMSD2.0 benchmark. We first reproduced the MV-CLIP model and confirmed a faithful ∼85% baseline. Through a quantitative and qualitative error analysis we identified a concrete weakness: the baseline over-predicts sarcasm because its additive fusion stage re-weights but does not allow the modalities to interact. To address this, we proposed CAMGF-CLIP, which augments MV-CLIP with adaptive gated fusion and bidirectional crossmodal attention, while keeping the rest of the architecture fixed so that the comparison isolates the effect of the fusion operator. On MMSD2.0, the reproduced baseline attains 85.4% accuracy under validation-based selection; gated fusion and cross-modal attention reach 83.5% and 84.7% respectively on the selected checkpoint. Although neither operator improves overall accuracy under this strict protocol, both reduce the precision–recall imbalance responsible for the baseline's overprediction of sarcasm, and cross-modal attention matches the baseline's accuracy with balanced precision and recall at its best epoch. Our analysis attributes the operators' lower selected-checkpoint scores to overfitting under the fixed training budget rather than to a failure of the mechanism. Future work includes incorporating an explicit contrastive incongruity objective, evaluating across multiple random seeds and additional benchmarks, and extending the analysis to cross-lingual and low-resource settings.

REFERENCES

[1.] L. Qin, S. Huang, Q. Chen, C. Cai, Y. Zhang, B. Liang, W. Che, and R. Xu, "MMSD2.0: Towards a reliable multi-modal sarcasm detection system," in *Findings of the ACL*, 2023, pp. 10834–10845.

[2.] Y. Cai, H. Cai, and X. Wan, "Multi-modal sarcasm detection in Twitter with hierarchical fusion model," in *Proc. ACL*, 2019, pp. 2506–2515.

[3.] R. Schifanella, P. de Juan, J. Tetreault, and L. Cao, "Detecting sarcasm in multimodal social platforms," in *Proc. ACM Multimedia*, 2016, pp. 1136–1145.

[4.] N. Xu, Z. Zeng, and W. Mao, "Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association," in *Proc. ACL*, 2020, pp. 3777–3786.

[5.] H. Pan, Z. Lin, P. Fu, Y. Qi, and W. Wang, "Modeling intra and intermodality incongruity for multi-modal sarcasm detection," in *Findings of the ACL: EMNLP*, 2020, pp. 1383–1392.

[6.] B. Liang, C. Lou, X. Li, L. Gui, M. Yang, and R. Xu, "Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs," in *Proc. ACM Multimedia*, 2021, pp. 4707–4715.

[7.] B. Liang, C. Lou, X. Li, M. Yang, L. Gui, Y. He, W. Pei, and R. Xu, "Multi-modal sarcasm detection via cross-modal graph convolutional network," in *Proc. ACL*, 2022, pp. 1767–1777.

[8.] H. Liu, W. Wang, and H. Li, "Towards multi-modal sarcasm detection via hierarchical congruity modeling

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

with knowledge enhancement," in *Proc. EMNLP*, 2022, pp. 4995–5006.
[9.] Y. Tian, N. Xu, R. Zhang, and W. Mao, "Dynamic routing transformer network for multimodal sarcasm detection," in *Proc. ACL*, 2023, pp. 2468–2480.
[10.] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML*, 2021, pp. 8748–8763.
[11.] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021.
[12.] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016, pp. 770–778.
[13.] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. ICML*, 2021, pp. 5583–5594.
[14.] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. NeurIPS*, 2021, pp. 9694–9705.
[15.] A. Vaswani *et al.*, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.
[16.] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.
[17.] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[18.] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. EMNLP*, 2014, pp. 1532–1543.
[19.] A. Joshi, P. Bhattacharyya, and M. J. Carman, "Automatic sarcasm detection: A survey," *ACM Computing Surveys*, vol. 50, no. 5, pp. 1–22, 2017.
[20.] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machineˇ learning: A survey and taxonomy," *IEEE TPAMI*, vol. 41, no. 2, pp. 423– 443, 2019.
[21.] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc. EMNLP*, 2017, pp. 1103–1114.
[22.] J. Arevalo, T. Solorio, M. Montes-y-Gomez, and F. A. Gonz´ alez, "Gated´ multimodal units for information fusion," in *Proc. ICLR Workshop*, 2017.
[23.] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. ACL*, 2019, pp. 6558–6569.
[24.] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *Proc. NeurIPS*, 2021, pp. 14200–14213.
[25.] J. Hessel and L. Lee, "Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!," in *Proc. EMNLP*, 2020, pp. 861–877.
[26.] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, 2019.
[27.] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
[28.] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *JMLR*, vol. 15, no. 1, pp. 1929–1958, 2014.
[29.] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.
[30.] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," *arXiv:1606.08415*, 2016.
[31.] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts, "Learning word vectors for sentiment analysis," in *Proc. ACL*, 2011, pp. 142–150.
[32.] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.
[33.] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP: System Demonstrations*, 2020, pp. 38–45.
[34.] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, 2019, pp. 8024–8035.
[35.] E. Cambria, B. Schuller, Y. Xia, and C. Havasi, "New avenues in opinion mining and sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 2, pp. 15–21, 2013.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

APPENDIX A

➢ *Per-Epoch Training Results*

For full transparency and to support reproducibility, we report the complete per-epoch training trajectory of all three models. These numbers underlie the curves in Fig. 11 and the selection analysis in Section VI.

Table 7 lists the validation accuracy at every epoch for the three models; the maximum of each column (in bold) is the checkpoint selected for the main results. Both proposed operators reach their maximum at epoch 2, whereas the baseline peaks at epoch 7.

Tables 8–10 give the test-set metrics recorded whenever a new validation maximum was reached during training (the only epochs at which the test set is evaluated). These tables make the early-peaking behaviour of the proposed operators explicit: both attain their strongest test scores at epoch 1, before the validation criterion settles on epoch 2.

Table 7 Validation Accuracy at Every Epoch. Bold Marks Each Model's Selected (Maximum-Validation) Checkpoint.

Epoch	Baseline	+ Gated	+ Cross-attn
0	0.8465	0.8398	0.8353
1	0.8412	0.8461	0.8427
2	0.8473	0.8519	0.8564
3	0.8482	0.8456	0.8390
4	0.8469	0.8498	0.8315
5	0.8491	0.8469	0.8340
6	0.8390	0.8461	0.8336
7	0.8539	0.8423	0.8328
8	0.8495	0.8423	0.8320
9	0.8508	0.8436	0.8336

Table 8 Baseline: Test Metrics at Improving-Validation Epochs.

Epoch	Acc.	Macro-F1	Macro-P	Macro-R
0	0.8525	0.8495	0.8467	0.8578
2	0.8582	0.8542	0.8514	0.8583
3	0.8406	0.8388	0.8390	0.8523
5	0.8507	0.8485	0.8468	0.8597
7	0.8538	0.8508	0.8478	0.8586

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

APPENDIX B

➢ *Implementation and Reproducibility Checklist*

We summarise the elements required to reproduce our results. The model uses the CLIP ViT-B/32 backbone with three Transformer fusion layers (hidden size 512, eight heads). Training uses AdamW with a learning rate of 5×10^{-4} for the new modules and 1×10^{-6} for CLIP, weight decay 0.05, batch size 32, gradient clipping at 5.0, dropout 0.1, a linear warm-up over 20% of steps, and ten epochs. The random seed is fixed, checkpoints are selected by maximum validation accuracy, and each model is stored under its own directory keyed by model name. The two proposed operators are dropin replacements for the baseline fusion and reuse all other components unchanged, so a single code base produces all three models by changing one argument. The CLIP weights are loaded from a local directory to allow fully offline execution.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

APPENDIX C

➢ *Computational Footprint*

Table 11 reports the wall-clock training time of each model in our environment (a single GPU, ten epochs). All three models are of comparable cost; the proposed operators add only a small overhead over the baseline, consistent with the complexity analysis in Section IV. The cross-modal attention variant in fact completed fastest in our run, as its attention modules introduce little overhead relative to the CLIP backbone and the per-epoch cost is dominated by the shared encoders and fusion layers. At inference the additional cost of either operator is negligible, so the proposed changes do not affect the deployability of the model.

Table 9 Gated Fusion: Test Metrics at Improving-Validation Epochs.

Epoch	Acc.	Macro-F1	Macro-P	Macro-R
0	0.8381	0.8377	0.8410	0.8473
1	0.8501	0.8492	0.8489	0.8556
2	0.8348	0.8344	0.8383	0.8444

Table 10 Cross-Modal Attention: Test Metrics at Improving-Validation Epochs.

Epoch	Acc.	Macro-F1	Macro-P	Macro-R
0	0.8381	0.8376	0.8398	0.8463
1	0.8547	0.8522	0.8514	0.8531
2	0.8472	0.8459	0.8448	0.8509

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

APPENDIX D

➢ *A Contrastive Incongruity Objective for Future Work*

Our error analysis indicates that the residual difficulty lies in distinguishing genuine image–text incongruity from agreement. A natural extension, which we leave to future work, is to add an auxiliary contrastive objective that explicitly shapes the embedding space so that incongruity becomes linearly separable. Concretely, let z_t and z_v denote the (normalised) pooled text and image embeddings of a post. For a batch of B posts, a symmetric contrastive loss in the style of CLIP encourages the model to align the two modalities of *matching* posts while separating mismatched ones:

$$\mathcal{L}_{\text{con}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(z_{t,i}^{\top} z_{v,i}/\tau)}{\sum_{j=1}^{B}\exp(z_{t,i}^{\top} z_{v,j}/\tau)}, \tag{15}$$

Where τ is a temperature. Crucially, for sarcasm one would not use this loss directly to align all pairs—since sarcastic posts are by definition incongruent—but rather as a *signal*: the degree of cross-modal (mis)alignment, $s_i = z_{t,i}^{\top} z_{v,i}$, becomes an explicit feature that the classifier can exploit, and the auxiliary loss is applied only to the non-sarcastic subset to calibrate what "agreement" looks like. The overall training objective would then be

$$\text{Ltotal} = \text{L} + \lambda\text{Lcon}, \tag{16}$$

With L the existing three-head cross-entropy of Eq. (14) and λ a weighting hyperparameter. Because the alignment score s_i depends jointly on both modalities, this objective is complementary to the cross-modal attention operator, which already provides the representations from which s_i is computed. Combining the two—cross-modal attention to build interaction-aware representations and a contrastive term to calibrate agreement—together with the stronger regularisation identified in Section VIII, is, in our view, the most promising direction for converting the mechanism's epoch-1 advantage into a reliably selected improvement.

Table 11 Wall-Clock Training Time (Ten Epochs, Single GPU).

Model	Training time
MV-CLIP (baseline)	3h21m
+ Gated fusion	3h18m
+ Cross-modal attn.	2h20m

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

APPENDIX E

➢ *Qualitative Error Categories*

Beyond the aggregate metrics, inspecting the prediction files by hand reveals several recurring categories of difficulty that help explain the residual errors and point towards targeted improvements. We describe the most salient ones here.

Consistent pairs misread as sarcastic. The dominant error type for the baseline, consistent with its over-prediction, occurs on posts whose caption and image genuinely agree but whose phrasing superficially resembles sarcasm (for example, enthusiastic captions paired with images that actually justify the enthusiasm). Here the additive fusion provides no mechanism to verify that the modalities agree, and the model defaults to the positive class.

World-knowledge-dependent sarcasm. A second category requires background knowledge not present in the image, such as awareness of a public event, a brand, or a cultural reference. Because neither the CLIP encoders nor the fusion layers have access to such external knowledge, these cases are hard for all three models and motivate the knowledgeenhanced approaches discussed in Section II.

Fine-grained region dependence. A third category depends on a small region of the image—a price tag, a piece of text within the photo, or a single object—being read against the caption. These cases are precisely where cross-modal attention has an advantage in principle, since its token-level attention can localise the relevant region, and they are a natural target for the region-aware extensions we leave to future work.

Ambiguous or subjective labels. Finally, a residual category comprises posts whose sarcasm is genuinely ambiguous even to a human reader; the subjectivity of sarcasm annotation places an upper bound on achievable accuracy and accounts for a portion of the errors that no architectural change is likely to resolve.

Together these categories suggest that future gains are most likely to come from (i) explicit modelling of cross-modal agreement, (ii) integration of external knowledge, and (iii) region-level grounding—each of which targets a distinct and identifiable slice of the current error distribution.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26jul207

APPENDIX F

- *Synthesis and Practical Recommendations*

Bringing together the quantitative results, the diagnostic metrics, the training dynamics and the qualitative analysis, three practical conclusions emerge for practitioners working on multimodal sarcasm detection with CLIP-based models.

First, the *fusion stage is a high-leverage and low-cost point of intervention*: meaningful changes in model behaviour (in particular the precision–recall balance) are obtainable by modifying only the operator that combines the two pooled representations, without retraining or replacing the expensive CLIP backbone. This makes fusion an attractive target for rapid experimentation.

Second, *capacity must be matched with regularisation*. Both proposed operators increase model capacity, and both reach their best generalisation very early in training before overfitting; under a fixed, frugal schedule this manifests as lower validation-selected scores. A practitioner adopting these operators should pair them with early stopping on a held-out set, a longer warm-up, or explicit regularisation, rather than a fixed epoch budget tuned for the lower-capacity baseline.

Third, *explicit interaction beats undirected capacity*. The contrast between the two operators is itself the clearest finding: cross-modal attention, which builds an explicit interaction between the modalities, tracks the baseline closely and improves positive-class precision, whereas gated fusion, which adds adaptive capacity without an interaction mechanism, amplifies the over-prediction. For a task whose very definition is cross-modal incongruity, this is the expected and encouraging outcome, and it motivates our recommendation to build future systems around explicit cross-modal interaction supplemented by the contrastive objective of Appendix D.