Token-Region Guided Cross-Attention Fusion for Multimodal Affect Interpretation

Musa Tur Farazi
Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology

Nufayer Jahan Reza
Department of Biomedical Engineering
Bangladesh University of Engineering and Technology

arXiv:2607.23493v1 [cs.CV] 26 Jul 2026

***Abstract*—Automated analysis of multimodal content on social networks has become a critical task for understanding public sentiment and information diffusion in the digital age. However, classifying internet memes remains computationally challenging due to the intricate interplay between visual cues and embedded, often stylized, text, particularly in low-resource languages like Bengali Language. This paper addresses the detection of political intent in Bengali memes by introducing Multimodal Cross-Attention Fusion framework. We first leverage a Vision-Language Model to extract high-fidelity OCR text from noisy meme images. Subsequently, we encode visual and textual features and synthesize them through a cross-modal multi-head attention mechanism that aligns semantic tokens with visual regions. We also investigate the integration of a domain-specific political lexicon as a knowledge prior. Experimental evaluation on the *PoliMemeDecode*[1] dataset shows that our attention-based fusion significantly outperforms unimodal baselines and standard concatenation methods, achieving a state-of-the-art Macro-F1 of approximately 0.94. Interpretability analyzes further confirm that the model effectively learns to ground textual semantics in visual evidence.**

I. Introduction

Internet memes have become a popular medium for sharing ideas and humor on social media. Although often lighthearted, they frequently carry serious political content, propaganda, or satire, particularly in the Global South. Detecting political intent in these memes is challenging due to the complex interplay between visual cues and embedded text, which often includes sarcasm, cultural references, and code-mixed languages such as Bengali. Most prior research has focused on high-resource languages, leaving a significant gap in automated understanding tools for low-resource languages.

To address this challenge, we present a practical multimodal pipeline designed to classify Bengali memes as either *Political* or *Non-Political*. As illustrated in Fig. 1, our approach operates in three integrated stages. First, we tackle the difficulty of reading stylized meme text using a Vision-Language Model (e.g. Qwen2-VL) for robust OCR extraction, which effectively handles noisy backgrounds and artistic fonts. Second, we encode the raw modalities using powerful pretrained backbones: CLIP extracts high-level visual embeddings, while XLM-RoBERTa captures multilingual textual semantics. Finally, these features are synthesized via our proposed Multimodal Attention Fusion (MAF) module. Unlike simple concatenation, MAF aligns visual and textual tokens through multi-head attention and incorporates a curated political lexicon as a domain prior to guide the classification process.

This work contributes to the field of low-resource affective computing by introducing a novel fusion architecture that specifically addresses the nuances of political meme detection. We demonstrate that moving beyond simple late fusion to a cross-modal attention mechanism significantly improves feature alignment between image and text. Furthermore, we show that integrating lightweight, lexicon-augmented knowledge acts as an effective domain prior, boosting performance in data-scarce regimes.

II. Related Work

Early meme classification research largely targeted English toxic/hateful content. The Hateful Memes Challenge demonstrated that unimodal models are insufficient and established multimodal baselines [1]. Subsequent work expanded to offensive meme detection, including GNN-based target identification [2] and multimodal frameworks such as MOMENTA [3], often relying on richer external resources in the same line as [4].

Recent English-centric advances move beyond simple fusion. **PromptHate** uses caption-based prompting with RoBERTa and reports strong AUROC/accuracy on HarM [5]. **ISSUES** applies textual inversion with CLIP and achieves competitive accuracy on the Hateful Memes unseen test set [6]. For interpretability, **Thakur et al.** combine CLIP and BERTweet features for example-based reasoning and show multimodal gains on MAMI [7]. **Hamza et al.** fine-tune VisualBERT for religiously hateful memes and improve performance over standard pretrained baselines [8].

In low-resource Bangla, progress is recent. MUTE provided an early Bengali hateful meme dataset [9]. Ahsan et al. introduced MIMOSA (including a Political category) and proposed attention-based fusion (MAF) [10]. Our work follows this direction but targets binary Political vs. Non-Political detection and adds a lexicon-based knowledge component.

A key challenge is combining the image and OCR text. Previous works often use late fusion via feature concatenation [1] [4] [3] [9], while newer methods model cross-modal interactions with co-/cross-attention, e.g., DORA [11] and attentive fusion [10]. We adopt multi-head attention to capture token–visual alignment, and we complement learned features with lightweight domain knowledge: a political lexicon signal (cf. graph-augmented approaches [2]).

[1]Dataset Available at : Kaggle/PoliMemeDecode

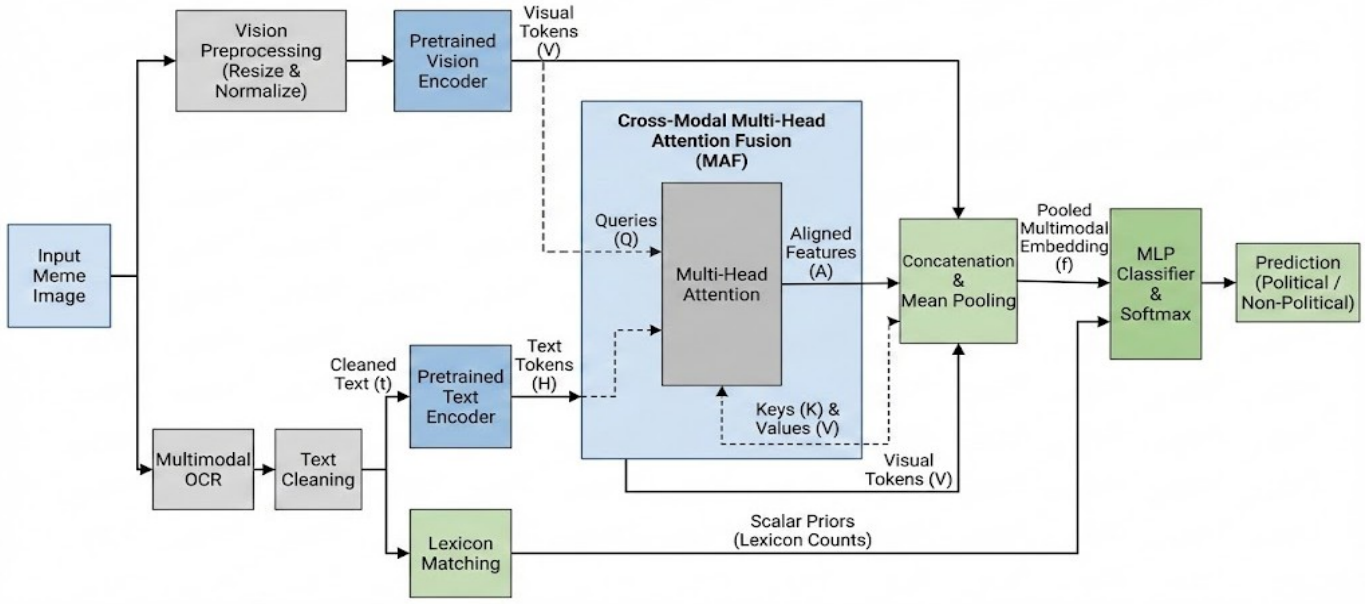


Fig. 1. The proposed Multimodal Attention Fusion (MAF) pipeline. The system extracts text using a VLM, encodes image and text via pretrained backbones, aligns them using cross-modal attention, and incorporates lexicon-based domain priors for final classification.

Finally, we leverage large pretrained models. CLIP provides strong visual representations [12], but Bengali OCR requires a multilingual text encoder; we pair CLIP with XLM-RoBERTa [13], similar in spirit to low-resource multimodal setups [11], while emphasizing tight attention-based fusion for Bengali political meme detection.

III. Problem Formulation and Evaluation Protocol

A. Task Definition

Let $\mathcal{D} = \{(x_i, \hat{t}_i, c_i, y_i)\}_{i=1}^{N}$, where x_i is an RGB meme image, $\hat{t}_i$ is cached OCR text, $c_i \in [0,1]$ is OCR confidence (set c_i=1 if unavailable), and $y_i \in \{0,1\}$ with 1=Political and 0=Non-Political. We learn a classifier producing logits $g_\theta(x, \hat{t}, c) \in \mathbb{R}^2$ and

$$p_\theta(y \mid x, \hat{t}, c) = \text{softmax}\big(g_\theta(x, \hat{t}, c)\big). \tag{1}$$

Prediction is $\hat{y} = \arg\max_{k\in\{0,1\}} p_\theta(y{=}k \mid x, \hat{t}, c)$.

B. Evaluation Metrics

We prioritize Macro-F1 and the Matthews Correlation Coefficient (MCC). For any class $k \in \{0,1\}$, the metrics are defined as:

$$\text{Prec}_k = \frac{TP_k}{TP_k + FP_k}, \qquad \text{Rec}_k = \frac{TP_k}{TP_k + FN_k} \tag{2}$$

$$F1_k = \frac{2 \cdot \text{Prec}_k \cdot \text{Rec}_k}{\text{Prec}_k + \text{Rec}_k}, \quad \text{MacroF1} = \tfrac{1}{2}\sum_k F1_k \tag{3}$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{4}$$

where TP, TN, FP, FN being calculated with respect to the positive class (Political) .

C. Cross-Validation

For fold k, $\mathcal{D}_{\text{train}}^{(k)} = \mathcal{D} \setminus \mathcal{D}_{\text{val}}^{(k)}$ with preserved class ratios. We evaluate out-of-fold (OOF) Macro-F1 over $\bigcup_k \mathcal{D}_{\text{val}}^{(k)}$. We ensemble K fold models by averaging probabilities:

$$\bar{p}(y \mid x, \hat{t}, c) = \frac{1}{K}\sum_{k=1}^{K} p_{\theta_k}(y \mid x, \hat{t}, c). \tag{5}$$

IV. Data Processing and Lexicon Prior

We use raw OCR output $\hat{t}$ and confidence c. We obtain clean text via

$$t = \textsc{Clean}(\hat{t}), \tag{6}$$

and resize and/or normalize images to the input resolution of the vision backbone (e.g., 224×224) with mild augmentation.

Algorithm 1 $\textsc{Clean}(\hat{t})$: OCR Text Normalization

Require: Raw OCR string $\hat{t}$
Require: Cleaned text t
1: $t \leftarrow$ lowercase($\hat{t}$)
2: Replace line breaks with spaces. Remove URLs and boilerplate. Keep Bengali Unicode with basic punctuation and drop other symbols.
3: Collapse repeated punctuation and whitespace
4: **return** t

We add a lexicon prior using a seed set $\mathcal{L}_0$ and a fold-only expansion. Lexicon matches are:

$$\ell(t) = \sum_{w\in\mathcal{L}} \mathbb{I}[w \subseteq t]. \tag{7}$$

Expansion uses smoothed log-odds:

$$s(w) = \log \frac{c_1(w) + \alpha}{C_1 + \alpha|\mathcal{V}|} - \log \frac{c_0(w) + \alpha}{C_0 + \alpha|\mathcal{V}|}. \tag{8}$$

Algorithm 2 Fold-Safe Lexicon Expansion

Require: Training texts/labels $\{(t_i, y_i)\}$, seed lexicon $\mathcal{L}_0$, smoothing α, top-K
Ensure: Expanded lexicon $\mathcal{L}$
1: Compute token counts $c_1(w)$ (Political) and $c_0(w)$ (Non-Political)
2: Compute $s(w)$ via Eq. (8) for $w \in \mathcal{V}$
3: $\mathcal{L}_{\text{auto}} \leftarrow$ top-K tokens with largest $s(w)$ and $s(w) > 0$
4: **return** $\mathcal{L} = \mathcal{L}_0 \cup \mathcal{L}_{\text{auto}}$

V. Lexicon-Augmented Multimodal Attention Fusion

Given an image x and cleaned OCR text t, we encode vision/text, align them with cross-modal attention, and append scalar priors $(\ell(t), c)$ for classification.

A. Encoders, Projection, and Fusion

CLIP yields visual tokens:

$$V = f_{\text{V}}(x) = [v_1, \ldots, v_M], \qquad v_j \in \mathbb{R}^{d_v}. \tag{9}$$

XLM-R yields text tokens:

$$H = f_{\text{T}}(t) = [h_1, \ldots, h_L], \qquad h_i \in \mathbb{R}^{d_t}. \tag{10}$$

Project to shared dimension d:

$$\tilde{v}_j = W_v v_j + b_v, \quad \tilde{h}_i = W_h h_i + b_h, \tag{11}$$

yielding $\tilde{V} \in \mathbb{R}^{M \times d}$ and $\tilde{H} \in \mathbb{R}^{L \times d}$.

B. Cross-Modal Multi-Head Attention

Queries come from text and keys/values from vision:

$$Q = \tilde{H}, \qquad K = \tilde{V}, \qquad V = \tilde{V}. \tag{12}$$

Scaled dot-product attention:

$$\text{Att}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right) V. \tag{13}$$

Multi-head attention:

$$\text{head}_r = \text{Att}(QW_r^Q, KW_r^K, VW_r^V), \tag{14}$$

$$A = \text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_R)W^O, \tag{15}$$

where $A \in \mathbb{R}^{L \times d}$. If only a global CLIP embedding is used ($M = 1$), replicate it:

$$\tilde{V}_{\text{rep}} = [\tilde{\tilde{v}}, \ldots, \tilde{\tilde{v}}] \in \mathbb{R}^{L \times d}. \tag{16}$$

C. Pooling, Scalars, and Classifier

Fuse and mean-pool:

$$F_i \;=\; [\tilde{h}_i \,;\, A_i \,;\, \tilde{v}_i] \in \mathbb{R}^{3d}, \qquad i = 1, \ldots, L, \tag{17}$$

$$\bar{f} \;=\; \frac{1}{L} \sum_{i=1}^{L} F_i \in \mathbb{R}^{3d}. \tag{18}$$

Append scalar priors:

$$s \;=\; [\,\phi(\ell(t)) \,;\, c\,] \in \mathbb{R}^2, \tag{19}$$

$$z \;=\; [\,\bar{f} \,;\, s\,] \in \mathbb{R}^{3d+2}. \tag{20}$$

Classifier logits:

$$u \;=\; \text{MLP}(\text{LayerNorm}(z)) \in \mathbb{R}^2. \tag{21}$$

Optional lexicon boosting for logit prior :

$$\tilde{u}_1 = u_1 + \beta\, \ell(t), \qquad \tilde{u}_0 = u_0. \tag{22}$$

Algorithm 3 Forward Pass of Lexicon-Augmented MAF

Require: Image x, raw OCR $\hat{t}$, confidence c
Ensure: Class probabilities $p \in [0, 1]^2$
$t \leftarrow \textsc{Clean}(\hat{t})$ ▷ Alg. 1
$\ell \leftarrow \sum_{w \in \mathcal{L}} \mathbb{I}[w \subseteq t]$ ▷ Eq. (7)
$V \leftarrow f_{\text{V}}(x)$, $H \leftarrow f_{\text{T}}(t)$
Project to shared dimension: $\tilde{V}, \tilde{H}$
if $M = 1$ **then**
    Replicate $\tilde{\tilde{v}}$ to length L to form $\tilde{V}_{\text{rep}}$
    $A \leftarrow \text{MHA}(Q = \tilde{H}, K = \tilde{V}_{\text{rep}}, V = \tilde{V}_{\text{rep}})$
else
    $A \leftarrow \text{MHA}(Q = \tilde{H}, K = \tilde{V}, V = \tilde{V})$
end if
$\bar{f} \leftarrow \frac{1}{L} \sum_{i=1}^{L} [\tilde{h}_i; A_i; \tilde{v}_i]$
$z \leftarrow [\bar{f}; \phi(\ell); c]$
$u \leftarrow \text{MLP}(\text{LayerNorm}(z))$
if lexicon boosting enabled **then**
    $\tilde{u} \leftarrow (u_0,\, u_1 + \beta\ell)$ ▷ Eq. (22)
else
    $\tilde{u} \leftarrow u$
end if
$p \leftarrow \text{softmax}(\tilde{u})$
return p

D. Training Objective and Optimization

1) Weighted Label-Smoothed Cross-Entropy: With $C = 2$ classes, class weights w_k, and smoothing $\varepsilon \in [0, 1)$:

$$\tilde{y}_{i,k} = (1 - \varepsilon)\, \mathbb{I}[k = y_i] + \frac{\varepsilon}{C}. \tag{23}$$

The objective:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{i=1}^{N} w_{y_i} \sum_{k=1}^{C} \tilde{y}_{i,k} \log p_{i,k}. \tag{24}$$

2) Optimization and Regularization: We optimize $\mathcal{L}(\theta)$ with AdamW, using cosine LR scheduling with warmup, dropout in the head, early stopping on validation Macro-F1.

3) Cross-Validation Training and Test-Time Ensembling: We train one model per stratified fold and compute OOF predictions; test predictions are ensembled by averaging fold probabilities.

Algorithm 4 Stratified K-Fold Training and Inference

Require: Dataset $\mathcal{D}$, number of folds K
Ensure: OOF predictions $\hat{y}^{\text{OOF}}$, test probabilities $\bar{p}^{\text{test}}$

```
Split D into stratified folds {(D_train^(k), D_val^(k))}_{k=1}^K
for k = 1 to K do
    Build fold-specific lexicon L^(k) using D_train^(k)
    Train model θ_k on D_train^(k) by minimizing Eq. (24)
    Predict on D_val^(k) to fill OOF predictions ŷ^OOF
    Predict probabilities on test set: p_test^(k)(·)
end for
p̄^test ← (1/K) Σ_{k=1}^K p_test^(k)
return ŷ^OOF, p̄^test
```

VI. Experiments

A. Experimental Setup

We evaluate PoliMemeDecode (Political vs. Non-Political) using **stratified 3-fold cross-validation**; all reported metrics are computed from aggregated **out-of-fold (OOF)** predictions.

Each meme provides an image and OCR text. We optionally include lexicon matches and OCR confidence as scalar features.

Vision uses OpenCLIP ViT-B/16 (LAION2B). Text uses XLM-RoBERTa-large with lower layers frozen (bottom 12 blocks),nd vision fine-tuning is enabled.

We evaluate ablations on image-only (A0), text-only (A1), text+scalars (A2), text+scalars+boost (A3), concat fusion (A4), MAF (A5), and MAF+scalars+boost (A6).

AdamW with warmup (ratio 0.1), LR 1×10^{-5}, weight decay 1×10^{-4}, early stopping (patience 1), and gradient accumulation. We report Macro-F1 and MCC, additionally ROC-AUC, AP, and Brier score.

B. Main Results

Table I reports mean $\pm$ std across folds. Both modalities are strong (A0 $>$ A1), while multimodal fusion improves performance; **MAF (A5)** is best. Lexicon boosting (A3 or A6) degrades performance, consistent with overconfident errors under keyword priors.

TABLE I
Stratified 3-fold CV results (mean $\pm$ std). Best overall is **A5 (MAF)**.

ID	Model Variant	Macro-F1	MCC
A0	Image Only (OpenCLIP ViT-B/16)	0.884 ± 0.016	0.770 ± 0.030
A1	Text Only (XLM-R large)	0.856 ± 0.006	0.713 ± 0.013
A2	Text + Scalars	0.861 ± 0.008	0.724 ± 0.014
A3	Text + Scalars + Boost	0.831 ± 0.005	0.663 ± 0.010
A4	Image + Text (Concat)	0.935 ± 0.011	0.871 ± 0.022
A5	**Image + Text (MAF)**	$\mathbf{0.940 \pm 0.005}$	$\mathbf{0.879 \pm 0.011}$
A6	Full (MAF + Scalars + Boost)	0.915 ± 0.010	0.831 ± 0.020

C. Comparison with Prior Multimodal Meme Models

Table II situates our results relative to representative multimodal meme classifiers on their *original* benchmarks (not directly comparable across datasets). Despite its simplicity, MAF achieves strong Macro-F1 on Bengali political memes without auxiliary captioning or inversion pipelines.

TABLE II
Performance landscape of representative multimodal meme classifiers (reported on their respective datasets; not directly comparable to PoliMemeDecode).

Method	Reference	Dataset	Metric	Score
Existing Baselines (English/Multimodal)				
VisualBERT	Hamza et al. (2024)	Religious Hate	Acc	70.60
Example-based	Thakur et al. (2023)	MAMI (Misogyny)	F1	0.701
ISSUES (CLIP+Inv)	Burbi et al. (2023)	Hateful Memes	Acc	77.70
PromptHate	Cao et al. (2022)	HarMeme	Acc	84.47
Proposed Method (Bengali)				
MAF (Ours)	This Work	**PoliMemeDecode**	**Macro-F1**	**0.940**

VII. Ablation Study and Analysis

We complement quantitative ablations with diagnostics to explain *why* attention-based fusion helps and *when* heuristics fail.

A. Impact of Modality and Fusion

Figure 2 summarizes Macro-F1 and MCC across variants; gains largely reflect simultaneous reductions in FP and FN rather than error trade-offs.

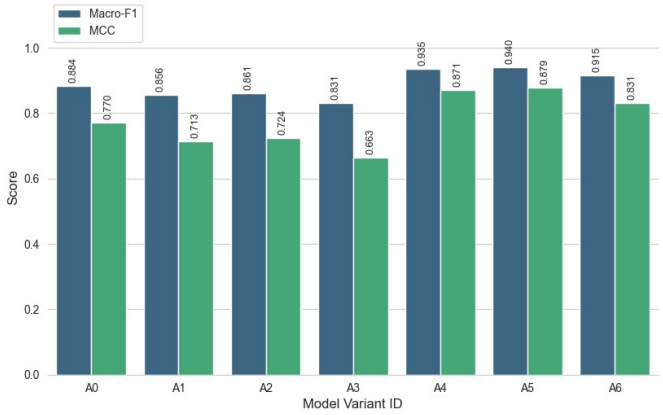


Fig. 2. Ablation performance (Macro-F1 and MCC). Multimodal fusion (A4/A5) yields large gains over unimodal baselines. MAF (A5) provides the strongest overall trade-off.

B. Confusion Matrix Analysis (Best Model: A5)

Figure 3 shows the confusion matrix of A5. With TN=1921, FP=86, FN=60, TP=793, political recall is $793/(793+60) = 0.930$ and precision is $793/(793+86) = 0.902$, indicating a strong balance of errors.

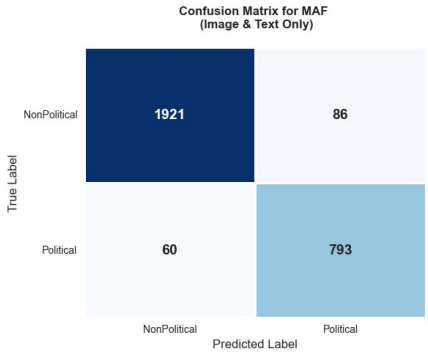


Fig. 3. Confusion matrix for A5 (MAF).

C. Representation Structure (t-SNE)

Figure 4 visualizes pooled features (fold-0 validation). Classes form largely separable regions with a narrow overlap; misclassifications cluster near this boundary, consistent with genuinely ambiguous memes.

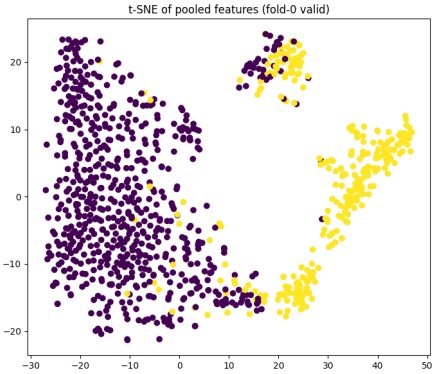


Fig. 4. t-SNE of pooled features (fold-0 validation). Political and non-political samples form largely separable regions with a small overlap band.

D. Calibration and Reliability

Figure 5 evaluates calibration for A5 (Brier=0.0419). The curve stays close to the diagonal, with only mild deviation.

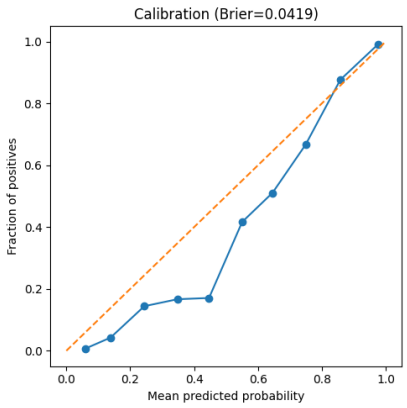
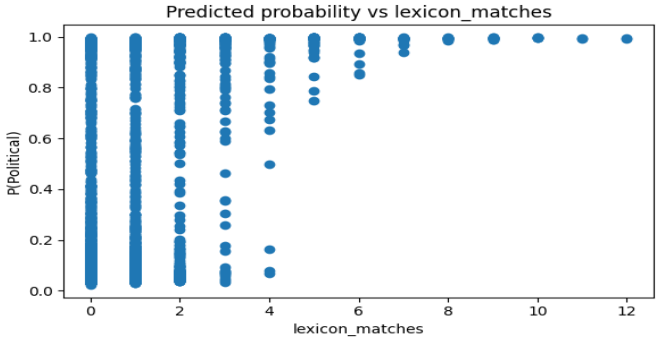


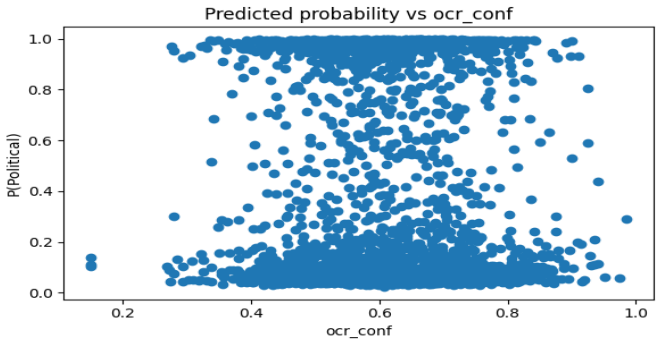
Fig. 5. Calibration curve for A5 (Brier=0.0419).

E. Role of Lexicon Matches and OCR Confidence

Figure 6 relates auxiliary scalars to model confidence. Lexicon matches correlates with higher political probability, but the wide spread at low counts shows predictions are not purely keyword-based.



(a) Predicted probability vs. lexicon matches



(b) Predicted probability vs. OCR confidence

Fig. 6. Relationship between auxiliary scalars and model confidence.

F. Visual Interpretability with SmoothGrad

Figure 7 shows SmoothGrad saliency: the model focuses on overlaid text and salient identity regions (e.g., faces/emblems) rather than backgrounds, supporting semantically grounded visual use.

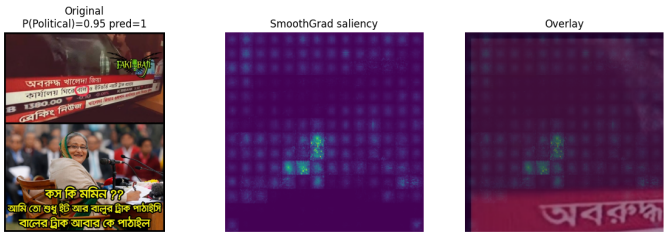


Fig. 7. SmoothGrad saliency (A5). Saliency concentrates on overlaid text and salient identity regions

G. OCR Token Contribution Analysis

We quantify OCR token influence by masking token w and measuring the expected probability drop:

$$\Delta(w) \;=\; \mathbb{E}\big[p(y{=}1 \mid x, t) \;-\; p(y{=}1 \mid x, t_{\setminus w})\big], \qquad (25)$$

where $t_{\setminus w}$ removes (or masks) w. Figure 8 reports the highest-magnitude positive/negative tokens. XLM-R uses subword units, so attributions are at the fragment level.

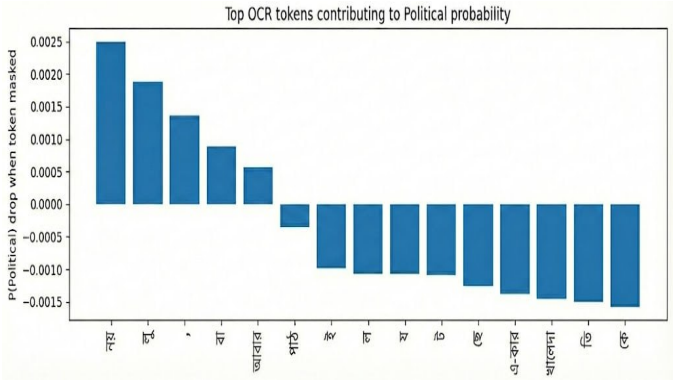


Fig. 8. Token-occlusion attribution on OCR text. Bars show the average change in predicted Political probability when a token is masked/removed. Positive values indicate tokens that increase Political probability; negative values indicate tokens that suppress it.

H. Why Lexicon Boosting Degraded Performance

Manual boosting adds a fixed prior term:

$$\mathbf{z}' = \mathbf{z} + \beta\, \ell(t), \tag{26}$$

which can increase recall when explicit political keywords appear, but also increases false positives in neutral or metaphorical contexts and worsens calibration relative to A5.

VIII. Discussion

a) Why multimodal attention helps.: MAF consistently improves over concatenation (A4) and strengthens Macro-F1/MCC, suggesting better alignment between salient OCR tokens and visual evidence, particularly for context-dependent or sarcastic memes.

b) Feature engineering vs. learned fusion.: Manual lexicon boosting (A3/A6) reduces performance and calibration, indicating brittle behavior when political keywords appear in neutral/metaphorical contexts. Learned fusion better conditions lexical cues on the image.

c) Role of OCR quality.: Accurate OCR remains important for stylized Bengali and code-mixed text; improved transcripts reduce ambiguity and enable stronger multimodal reasoning.

IX. Conclusion

We studied Bengali political meme detection with a multimodal framework that jointly leverages meme imagery and OCR text. Our proposed **Multimodal Attention Fusion (MAF)** aligns token-level textual evidence with visual cues via cross-modal multi-head attention, and we further explored lightweight domain signals through a political lexicon prior. On the PoliMemeDecode dataset, MAF consistently outperformed unimodal baselines and simple concatenation, achieving the best overall performance with **Macro-F1** ≈ 0.94 and **MCC** ≈ 0.88 under stratified cross-validation. Extensive analyses (confusion matrix, calibration, t-SNE, saliency, and token-occlusion) indicate that the model learns complementary cross-modal evidence rather than relying on brittle keyword triggers, and they explain why manual lexicon boosting can harm calibration and increase false positives. Overall, our results highlight that attention-based fusion of strong pretrained backbones (OpenCLIP + XLM-R) is an effective and practical approach for low-resource, script-diverse political meme understanding.

References

[1] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Advances in neural information processing systems*, vol. 33, pp. 2611–2624, 2020.

[2] L. Shang, C. Youn, Y. Zha, Y. Zhang, and D. Wang, "Knowmeme: A knowledge-enriched graph neural network solution to offensive meme detection," in *2021 IEEE 17th International Conference on eScience (eScience)*, 2021, pp. 186–195.

[3] S. Pramanick, S. Sharma, D. Dimitrov, M. S. Akhtar, P. Nakov, and T. Chakraborty, "MOMENTA: A multimodal framework for detecting harmful memes and their targets," in *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 4439–4455. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.379/

[4] S. Sharma, M. S. Akhtar, P. Nakov, and T. Chakraborty, "Disarm: Detecting the victims targeted by harmful memes," *arXiv preprint arXiv:2205.05738*, 2022.

[5] R. Cao *et al.*, "Prompting for multimodal hateful meme classification," in *EMNLP*, 2022.

[6] G. Burbi *et al.*, "Mapping memes to words for multimodal hateful meme classification," in *ICCVW*, 2023.

[7] A. K. Thakur *et al.*, "Multimodal and explainable internet meme classification," *arXiv preprint arXiv:2212.05612*, 2023.

[8] A. Hamza *et al.*, "Multimodal religiously hateful social media memes classification," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 2024.

[9] E. Hossain, O. Sharif, and M. M. Hoque, "MUTE: A multimodal dataset for detecting hateful memes," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, Y. Hanqi, Y. Zonghan, S. Ruder, and W. Xiaojun, Eds. Online: Association for Computational Linguistics, Nov. 2022, pp. 32–39. [Online]. Available: https://aclanthology.org/2022.aacl-srw.5/

[10] S. Ahsan, E. Hossain, O. Sharif, A. Das, M. M. Hoque, and M. Dewan, "A multimodal framework to detect target aware aggression in memes," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Y. Graham and M. Purver, Eds. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2487–2500. [Online]. Available: https://aclanthology.org/2024.eacl-long.153/

[11] E. Hossain, O. Sharif, M. M. Hoque, and S. M. Preum, "Deciphering hate: identifying hateful memes and their targets," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 8347–8359.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021, pp. 8748–8763.

[13] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 8440–8451. [Online]. Available: https://aclanthology.org/2020.acl-main.747/