KEC@MMLow 2026: Multimodal Fusion for Hierarchical Tamil Political Meme Classification

Asvitha RM and Amirthamozhi VS
Department of Computer Science and Engineering
Kongu Engineering College
Erode, Tamil Nadu, India

Abstract

Political memes are inherently multimodal, combining visual cues, embedded text, and code-mixed language to express political stance and target specific political entities. This makes automated analysis particularly challenging in low-resource settings such as Tamil. We present the KEC system for the MMLow 2026 shared task at AACL-IJCNLP 2026, which requires hierarchical classification of Tamil political memes at two levels: stance identification and political target identification. Our approach combines XLM-RoBERTa and Vision Transformer representations with OCR-derived textual information, multimodal fusion, and bidirectional cross-modal attention. We further introduce hierarchy-aware prediction in which Level-2 target identification is conditioned on the Level-1 stance distribution, together with class-weighted objectives, adaptive task weighting, and hierarchical consistency regularization. A controlled six-stage ablation study and three-seed stability analysis are conducted on the validation split. The final hierarchical model obtains a Level-1 macro-F1 of $78.77 \pm 1.21\%$ and Level-2 macro-F1 of $46.07 \pm 3.23\%$, with a joint exact-match accuracy of $55.40 \pm 2.34\%$ across seeds 42, 123, and 2024. Our analysis indicates that visual information is particularly beneficial for stance classification, whereas multimodal fusion plays an important role in target identification, while OCR augmentation and hierarchical conditioning introduce performance trade-offs at Level 2.

1 Introduction

Political memes have become an important medium for expressing opinions, satire, support, and opposition on social media. Unlike conventional text-based content, memes communicate through interactions among visual elements, written text, symbols, and context-dependent cues. This multimodal nature makes interpretation challenging, particularly in low-resource and code-mixed languages such as Tamil, where memes frequently combine Tamil, English, Romanized Tamil, informal spellings, and text embedded within images.

The MMLow 2026 shared task at AACL-IJCNLP 2026 (Rajiakodi et al., 2026) addresses this challenge through hierarchical classification of Tamil political memes. The task requires systems to first determine the stance expressed by a meme at **Level 1** (*Support/Praise* or *Troll/Oppose*) and then identify its political target at **Level 2**. These decisions are closely coupled, as target interpretation can depend on stance polarity, while textual and visual evidence may contribute differently to each level.

In this study, team **KEC** presents a multimodal framework combining XLM-RoBERTa for code-mixed text, a Vision Transformer for visual information, and OCR-derived text extracted from meme images. We investigate both feature concatenation and bidirectional cross-modal attention, adapting the pretrained encoders using **Low-Rank Adaptation (LoRA)**. The final model incorporates hierarchical conditioning by using the Level-1 stance distribution for Level-2 prediction and jointly optimizes both objectives with class balancing, **homoscedastic uncertainty weighting**, and hierarchical consistency regularization. Our controlled experiments further examine the effects of visual information, OCR augmentation, multimodal fusion, cross-modal attention, and hierarchical conditioning through a six-stage ablation and three-seed stability analysis. [1]

2 Task and Dataset

The MMLow 2026 shared task performs hierarchical classification of Tamil political memes. Level 1 classifies stance as *Support/Praise* or *Troll/Oppose*, while Level 2 identifies the political target. The official Tamil dataset (Rajiakodi et al., 2026) con-

[1] Code is available on GitHub.

Proceedings of the Fifth Workshop on Multimodal Machine Learning in Low-resource Languages, pages 86–90
November 9, 2026 ©2026 Association for Computational Linguistics

tains 969 training samples and 244 unlabeled test samples.

2.1 Data Preparation

A data-quality audit excluded five hierarchy-inconsistent samples, leaving 964 instances; six caption-missing samples were retained using OCR fallback. The cleaned data were split into 819 training and 145 validation samples using an 85:15 stratified split on joint Level-1/Level-2 labels, with text normalized using Unicode NFC.

2.2 Evaluation

The shared task evaluates classification performance using macro-averaged precision, recall, and F1 score across the classes. Accordingly, we use macro-F1 as the primary metric for evaluating the two classification levels, while also reporting the joint exact-match accuracy for the hierarchical prediction.

3 Methodology

Our system is a configurable multimodal architecture that combines textual, visual, and OCR-derived information for hierarchical meme classification. The same architecture is used throughout a controlled six-stage ablation, with components introduced progressively to isolate their contributions.

3.1 Text and OCR Representation

We use XLM-RoBERTa-base (Conneau et al., 2020) to encode the Tamil–English code-mixed textual content. Input sequences are truncated or padded to a maximum length of 160 tokens. Captions are Unicode-normalized using NFC normalization before tokenization.

Since important information may be embedded in meme images rather than captions, we use OCR as an auxiliary textual source, particularly for names, slogans, and target-related text. We employ EasyOCR (JaidedAI, 2023) with Tamil and English models, discarding detections below 0.4 confidence and removing consecutive duplicates. OCR text is combined with the normalized caption using a `[SEP]` marker, while caption-missing samples use OCR alone. Although OCR can introduce noise from stylized, low-resolution, and code-mixed text, retaining it allows us to quantify this trade-off explicitly in our ablation study.

To adapt the pretrained language encoder efficiently to the small training set, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) to the query and value projections of its attention layers, using rank $r = 8$, scaling factor $\alpha = 16$, and dropout 0.1. Across the two pretrained encoders, LoRA leaves approximately 1.09% of the total model parameters trainable, reducing the number of updated parameters while retaining task-specific adaptation.

3.2 Visual Representation

Visual features are extracted using a pretrained ViT-Base/16 encoder (Dosovitskiy et al., 2021). Images are converted to RGB and resized to 224×224 using aspect-ratio-preserving resize-with-padding, thereby avoiding distortion of faces, logos, political symbols, and embedded text. During training, light brightness and contrast perturbations and a small random crop are applied; validation uses deterministic resizing and normalization.

The ViT encoder is also adapted using LoRA with $r = 8$, $\alpha = 16$, and dropout 0.1, targeting its transformer qkv projections.

3.3 Multimodal Fusion

We evaluate direct feature concatenation and bidirectional cross-modal attention based on multi-head attention (Vaswani et al., 2017). For cross-attention, text and image tokens are projected into a shared 512-dimensional space and processed by separate 8-head attention modules in both directions. Each path uses residual connections and layer normalization, followed by mean pooling and concatenation.

3.4 Hierarchical Prediction

The fused representation is passed to separate prediction heads for the two task levels. The Level-1 head uses a 256-dimensional hidden layer with GELU activation and dropout of 0.2 to predict stance. In the final hierarchical configuration, the resulting Level-1 probability distribution is concatenated with the fused representation and supplied to the Level-2 head:

$$p_1 = softmax(W_1 h + b_1), \qquad h_2 = [h; p_1]. \tag{1}$$

The Level-2 head uses the same 256-dimensional hidden layer, GELU activation, and dropout to predict the political target.

3.5 Hierarchical Training Objective

Both levels use class-weighted cross-entropy, with class weights computed from the training partition. For the final hierarchical model, the two task losses are adaptively weighted using homoscedastic uncertainty (Kendall et al., 2018). Let $\mathcal{L}_1$ and $\mathcal{L}_2$

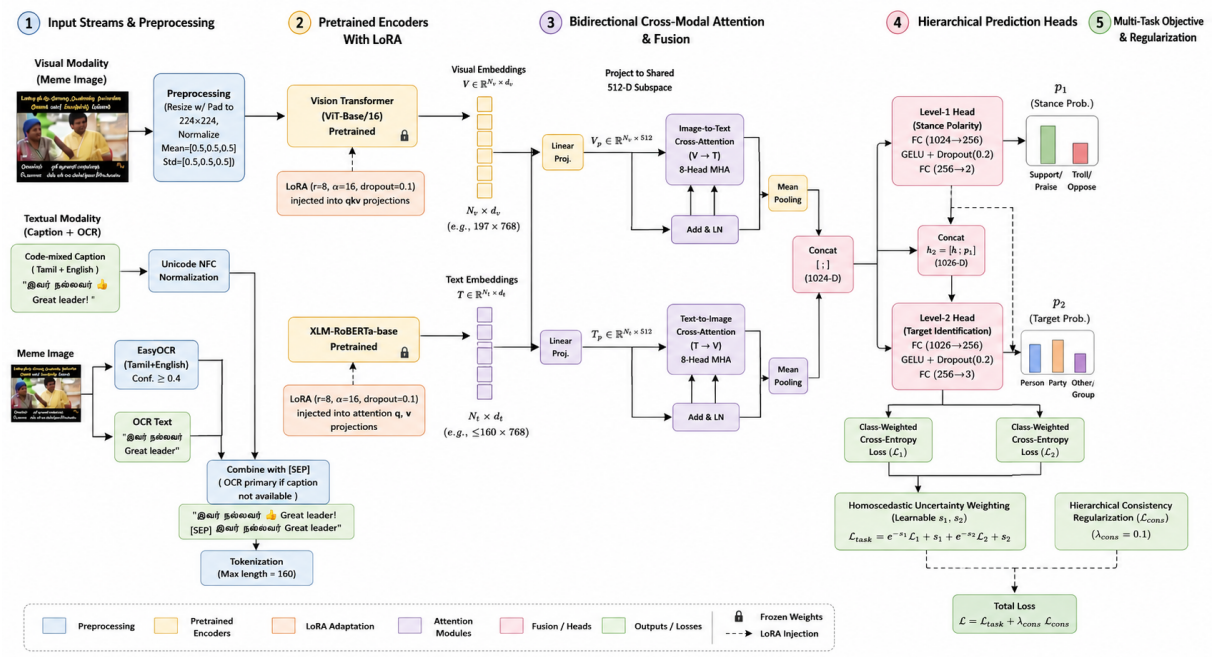


Figure 1: Overview of the proposed hierarchy-aware multimodal framework, combining LoRA-adapted XLM-RoBERTa and ViT-Base/16, bidirectional cross-modal attention, hierarchical conditioning, and uncertainty-weighted multi-task learning.

denote the Level-1 and Level-2 losses, and s_1, s_2 their learnable log-variance parameters. The task objective is

$$\mathcal{L}_{\text{task}} = e^{-s_1}\mathcal{L}_1 + s_1 + e^{-s_2}\mathcal{L}_2 + s_2. \quad (2)$$

We additionally apply a hierarchical consistency penalty to discourage Level-2 probability mass from being assigned to targets incompatible with the corresponding Level-1 class. The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{cons}}\mathcal{L}_{\text{cons}}, \quad (3)$$

where $\lambda_{\text{cons}} = 0.1$.

3.6 Optimization

All experiments use AdamW (Loshchilov and Hutter, 2019) (2×10^{-5} learning rate, 0.01 weight decay) for up to 12 epochs, with 10% warm-up, cosine decay, gradient clipping at 1.0, and mixed precision. Early stopping monitors validation Level-2 macro-F1 with patience 4; the best checkpoint is restored before evaluation. Because selection is based on Level-2 performance, it does not necessarily maximize Level-1 performance, which is reported separately.

4 Experiments and Results

4.1 Experimental Setup

Experiments use the cleaned 819/145 training–validation split with seed 42. We evaluate six progressively modified configurations from unimodal models through multimodal fusion, OCR, cross-modal attention, and hierarchical prediction. Macro-F1 is the primary metric at both levels; joint exact-match accuracy is also reported.

4.2 Ablation Study

Configuration	L1 Macro-F1	L2 Macro-F1
A – Text-only	0.2968	0.3015
B – Image-only	0.7654	0.3325
C – Multimodal Concatenation	0.7384	**0.5181**
D – Multimodal + OCR	0.7486	0.4458
E – Cross-attention + OCR	0.7851	0.3889
F – Final Hierarchical	**0.7860**	0.4448

Table 1: Single-run ablation results on the validation set using seed 42. Macro-F1 is reported separately for Level 1 and Level 2.

We conducted a six-stage ablation on the same validation split with seed 42, progressively adding visual information, multimodal concatenation, OCR, bidirectional cross-modal attention, and hierarchical conditioning with uncertainty-weighted training. Configuration F represents the complete hierarchy-aware system, rather than the configuration with the highest Level-2 score: C achieves the best single-run Level-2 macro-F1 (0.5181), whereas F achieves the best Level-1 macro-F1 (0.7860).

Visual information benefits Level 1, with macro-F1 increasing from 0.2968 (text-only) to 0.7654

Evaluation	Level-1 Macro-F1	Level-2 Macro-F1	Joint/Final
Our 3-seed validation	$78.77 \pm 1.21\%$	$46.07 \pm 3.23\%$	$55.40 \pm 2.34\%$
Official shared-task	—	0.4329	0.4329

Table 2: Internal three-seed validation results and official shared-task evaluation for KEC Run 1.

(image-only) and 0.7860 (F). For Level 2, concatenation achieves 0.5181, while OCR and cross-modal attention reduce performance to 0.4458 and 0.3889; hierarchical conditioning reaches 0.4448. The cleaned dataset contains 690 Troll/Oppose and 274 Support/Praise samples at Level 1, and 544 Troll/Oppose Against Person, 176 Support for person, 146 Troll/Oppose Against Party, and 98 Support for party at Level 2. Six samples lack captions; the text-only configuration uses the original text without OCR. Its macro-F1 of 0.3015 exceeds the Level-2 majority-class reference of 0.1806.

We also evaluated the consistency penalty with the same architecture, split, and seed. With $\lambda_{\text{cons}} = 0.1$, Level-1 macro-F1, Level-2 macro-F1, and joint exact-match accuracy were 0.7860, 0.4448, and 0.5793; without the penalty, they were 0.7823, 0.4176, and 0.5379, respectively. Paired permutation tests (10,000 permutations) and paired-bootstrap 95% confidence intervals on the 145-sample split gave a Level-1 difference of 0.0037 (CI: [-0.0204, 0.0326], $p = 1.0000$) and a Level-2 difference of 0.0272 (CI: [-0.0410, 0.0955], $p = 0.4396$). An exact McNemar test for joint exact-match accuracy gave $p = 0.3269$. These results do not establish statistically significant improvements from the consistency penalty on this validation split.

4.3 Multi-Seed Stability

We evaluated the final hierarchical model with seeds 42, 123, and 2024. Table 2 reports the mean and standard deviation alongside the official KEC Run 1 evaluation, which reports Level-2 score 0.4329 but no Level-1 score. The submission ranked 7th.

5 Discussion and Error Analysis

5.1 Main Findings

Stance classification benefits primarily from visual cues, whereas target identification requires complementary textual and visual evidence. The weaker Level-2 results with OCR and cross-modal attention suggest that additional multimodal processing can introduce noise in this low-resource setting, while greater seed variability indicates higher sensitivity to initialization.

5.2 Error Analysis

Fine-grained target identification remains challenging due to class imbalance and ambiguous textual–visual evidence. Person–party ambiguity is recurring: a meme may display a politician while targeting the associated party. OCR also adds noise for stylized, low-resolution, and code-mixed text, reducing Level-2 macro-F1 from 0.5181 to 0.4458.

We analyze the final model on the 145-sample validation split; Table 3 reports class-wise Level-2 precision, recall, and F1.

Level-2 Class	Precision	Recall	F1
Support for party	0.500	0.133	0.211
Support for person	0.484	0.577	0.526
Troll/Oppose Against Party	0.292	0.318	0.304
Troll/Oppose Against Person	0.721	0.756	0.738

Table 3: Class-wise Level-2 precision, recall, and F1 on the validation split.

Troll/Oppose Against Person has the highest F1 (0.738), while Support for party is most difficult (0.211). Table 4 shows confusion among support classes and between the two Troll/Oppose targets.

True / Pred.	Support party	Support person	Troll party	Troll person
Support for party	2	8	2	3
Support for person	2	15	2	7
Troll/Oppose Against Party	0	1	7	14
Troll/Oppose Against Person	0	7	13	62

Table 4: Level-2 confusion matrix on the 145-sample validation split.

6 Limitations and Conclusion

6.1 Limitations

Our study is constrained by the small, imbalanced training set and the noise introduced by OCR on stylized, low-resolution and code-mixed text. Validation results may also not generalize to memes from different periods, platforms or linguistic settings.

6.2 Conclusion

We presented the KEC system for the MMLow 2026 shared task, combining multilingual text, visual features, OCR, multimodal fusion, and hierarchy-aware learning for Tamil political meme classification. Visual information was most effective for Level-1 stance prediction, while direct multimodal concatenation gave the strongest Level-2 target identification. The multi-seed results showed stable stance performance but greater variability for target identification, highlighting the challenges of multimodal learning in low-resource, code-mixed political meme data.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7482–7491.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, B Premjith, CN Subalalitha, Rahul Ponnusamy, Anshid K A, Bhuvaneswari Sivagnanam, V Jananayagan, N Ragavan, P Santhini, and Bharathi Raja Chakravarthi. 2026. Overview of the shared task on multilevel political meme classification in tamil and malayalam. In *Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

A Additional Implementation Details

The experiments use XLM-RoBERTa-base and ViT-Base/16 with LoRA adaptation. For both encoders, LoRA uses rank $r = 8$, scaling factor $\alpha = 16$, and dropout 0.1. Text sequences are limited to 160 tokens, while meme images are resized to 224×224 using aspect-ratio-preserving padding with bilinear interpolation.

OCR is performed using EasyOCR with Tamil and English recognition models, discarding detections with confidence below 0.4. Cross-modal attention uses 512-dimensional projections and eight attention heads. Both prediction heads use a 256-dimensional hidden layer with GELU activation and dropout 0.2.

Training uses AdamW with learning rate 2×10^{-5} and weight decay 0.01, for up to 12 epochs with 10% warm-up and cosine decay. Training and validation use a batch size of 16, with 145 samples in the validation split. Gradient norms are clipped to 1.0, mixed-precision training is enabled, and early stopping monitors validation Level-2 macro-F1 with patience four. The controlled ablations use random seed 42, while the stability analysis uses seeds 42, 123, and 2024. Experiments were run on an NVIDIA GeForce RTX 3050 Laptop GPU; the six ablation runs required approximately 61 minutes in total.