Projukti Kavalan@MMLow 2026: Weighted Visual Frame and Cross-modal Fusion for Misogynistic Video Detection in Dravidian Language

Palaniraj I*
BestWay Solution
Tamil Nadu, India
pazhaniraj@bestwaysolution.in

Dwaipayan Dutta*
Dept. of Mathematics
Indian Institute of Space Science and Technology
Kerala, India
dwaipayandutta.25@res.iist.ac.in

Abstract

Misogyny and hate content have become a serious issue in the current digital era. Such content predominantly manifests as short video clips, making automated detection mechanisms imperative for social media platforms. To effectively classify such videos, a multimodal setup is required, as video data contains image frames, audio clips, and text captions. The multimodal misogyny video classification task aims to combine all audio, visual, and extracted text information. Conventional methods typically aim to process this extracted information in parallel to reach a conclusion. Our proposed framework utilizes the pre-trained models DINOv2, wav2vec 2.0, and MuRiL-large to embed visual, audio, and text features, respectively. To mitigate overfitting on datasets of different sizes, we propose a dual-architecture approach: we utilize a regularized Gated Early-Fusion Network (EF-Net) for the highly constrained Tamil dataset, and a deeper Cross-Modal Temporal Attention Network (TA-Net) with an ensemble strategy for the Malayalam dataset. Ultimately, our framework achieved a 68.8% macro-F1 score for the Malayalam dataset and an 80.7% macro-F1 score for the Tamil dataset, securing Rank 1 in misogyny video classification for both languages.

1 Introduction

Social media platforms have amplified gender-biased hate, increasingly distributed via multimodal formats like short videos. While robust detection architectures exist for high-resource languages, models for Dravidian languages like Tamil and Malayalam—which have massive digital footprints—lag significantly behind, necessitating specialized multimodal approaches capable of identifying misogyny across speech, text, and visual streams.

2 Related Work

Early research in misogyny detection concentrated on unimodal text models (Cao et al., 2020). As harmful content migrated to multimedia, the community adopted multimodal strategies, such as bimodal fusion for images and captions (Jindal et al., 2024) or chain-of-thought reasoning for memes (Kumari et al., 2024).

In video-based detection, filtering noisy temporal streams is paramount. Wang et al. (2023) introduced a frame-weighting mechanism utilizing cross-modal attention to isolate informative visual frames against on-screen text, which Céspedes-Sarrias et al. (2025) expanded using audio. While Céspedes-Sarrias et al. (2025) relies on standard cross-modal attention across parallel streams, our framework directly regulates the visual sequence. We adapt the frame-weighting concept from Wang et al. (2023) but significantly extend it: instead of their additive fusion for a single semantic stream, we introduce an independent audio weight vector (w_{audio}) and employ a strict multiplicative consensus mechanism ($\odot$). This ensures video frames are emphasized only upon simultaneous audio, text, and visual agreement, drastically reducing noise in constrained datasets.

Crucially, while these models excel on massive datasets, they suffer catastrophic overfitting on low-resource languages. Recent shared tasks (Chakravarthi et al., 2021; G et al., 2025; Rajiakodi et al., 2025) emphasize that multimodal hate speech detection in Dravidian languages remains a critical challenge. Our framework addresses this gap by strictly constraining architecture capacity based on dataset size for the MMLow 2026 shared task.

3 Methodology

Given the extreme resource differences between the Tamil and Malayalam datasets, a "one-size-fits-all" model is sub-optimal. Instead, we extract robust tri-

Proceedings of the Fifth Workshop on Multimodal Machine Learning in Low-resource Languages, pages 76–85
November 9, 2026 ©2026 Association for Computational Linguistics

modal features and route them to specialized architectures: a highly-regularized Gated Early-Fusion Network (**Architecture A**) for the extremely constrained Tamil dataset, and a deeper Cross-Modal Temporal Attention Network (**Architecture B**) for the larger, balanced Malayalam dataset.

3.1 Data Preprocessing and Feature Extraction

Two different video datasets are being used for this work: 146 total videos for Tamil (120 train, 26 test) and 343 for Malayalam (303 train, 40 test). Raw videos were segmented into visual frames (Culjak et al., 2012), audio was separated into vocal tracks (Rouard et al., 2023), and transcripts were generated via Gemini Flash 3.5 (Team et al., 2023). We extracted 1024-dimensional embeddings using DINOv2 (Oquab et al., 2023) for visuals, MuRiL-Large (Khanuja et al., 2021) for text, and wav2vec 2.0 (Baevski et al., 2020) for audio.

3.2 Architecture A: Gated Early-Fusion Network

To prevent memorization on the highly constrained Tamil dataset, we designed a Gated Early-Fusion Network operating on globally pooled features (F_v, F_a, F_t). Collapsing temporal sequences creates a strong informational bottleneck (regularizer). To suppress noise, we apply a learnable gate: $f_m = \mathrm{ReLU}(W_f F_m + b_f)$, $g_m = \sigma(W_g F_m + b_g)$, $\hat{F}_m = \mathrm{Dropout}(f_m \odot g_m)$, where $W_f \in \mathbb{R}^{64\times1024}$ compresses raw embeddings to 64 dimensions. Surviving features are concatenated $F_{fused} = [\hat{F}_v; \hat{F}_t; \hat{F}_a]$ and passed through a Residual MLP.

High-Capacity Variant: `EF-Net (High-Cap)` replaces the 64-dim bottleneck with a 512-dim projection (1536-dim concatenated) and a deeper MLP (0.4 dropout) to serve as a capacity-matched baseline counterpart.

3.3 Architecture B: Cross-Modal Temporal Attention Network

For the more balanced and contextually complex Malayalam dataset, we implemented a temporal attention architecture consisting of a Multimodal Information Encoding Module (MIE), a Visual Frame Weighting Module (VFW), and a Cross-Modal Transformer Fusion network (CMTF).

3.3.1 Problem Formulation

Given dataset $\mathcal{D} = \{(\mathcal{V}_i, y_i)\}$, with $y_i \in \{0, 1\}$, each video contains three modalities: visual $V \in \mathbb{R}^{n\times d_v}$, audio $A \in \mathbb{R}^{8\times d_a}$, and text $T \in \mathbb{R}^{1\times d_t}$. To maintain strict temporal alignment, visual streams are uniformly interpolated to 30 frames, and audio to 8 steps. Transcripts are represented as a globally pooled vector.

3.3.2 Multimodal Information Encoding Module (MIE)

Each representation is projected to a short dimension $d(= 64)$. The visual and audio representations are passed through a bidirectional GRU (Gated Recurrent Unit) model to encode information in both directions. Therefore visual encoder produces $F_v \in \mathbb{R}^{n\times d}$, audio encoder produces $F_a \in \mathbb{R}^{8\times d}$. For each video, we select the reference training clip with the same label whose visual representation has the highest cosine similarity with the video:

$$\mathcal{V}^+ = \{\mathcal{V}_{ref} \mid \arg\max_{V_{ref}} \mathrm{sim}(V, V_{ref}),\ y_V = y_{V_{ref}}\}$$

where V is the visual representation of the video clip $\mathcal{V}$, and V_{ref} is the visual representation of the reference video clip $\mathcal{V}_{ref}$. These positive visual references are encoded by the visual encoder and produce $F_v^+ \in \mathbb{R}^{n\times d}$. Moreover, the transcript representations pass through a linear projection and layer normalization layer, and so, the transcript encoder produces $F_t \in \mathbb{R}^{1\times d}$. Final encodings are: $F_v = \mathrm{Vis}(V)$, $F_v^+ = \mathrm{Vis}(V_{ref})$, $F_a = \mathrm{Aud}(A)$, and $F_t = \mathrm{Txt}(T)$.

3.3.3 Visual Frame Weighting Module (VFW)

The core of this architecture adapts a multimodal frame-weighting mechanism originally proposed by (Wang et al., 2023). From their foundational framework, we reuse the visual self-weight (w_{self}), the text-based cross-modal weight (w_{text}), and the contrastive reference-context weight (w_{ctx}) derived from positive sample alignment. To elevate this to a robust trimodal architecture, we introduce a novel cross-modal video-audio weight vector (w_{audio}). Furthermore, rather than using an additive fusion as in (Wang et al., 2023), we propose a multiplicative consensus fusion via the Hadamard product ($\odot$). Prior to this operation, any multi-step temporal sequences (such as audio $F_a \in \mathbb{R}^{8\times d}$ or visual $F_v \in \mathbb{R}^{n\times d}$) serving as the context c are first reduced to a single vector $c \in \mathbb{R}^d$ via an attention-based pooling layer. This pooling operation produces a unified context representation for subse-

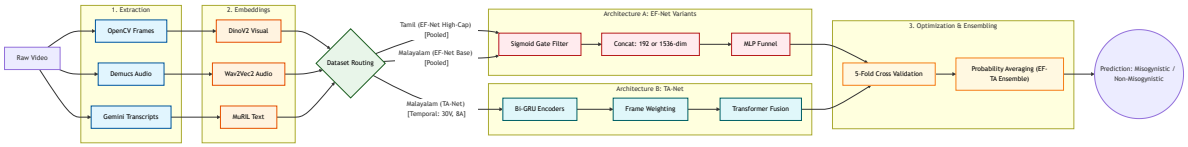


Figure 1: Overview of the proposed pipeline: text, visual, and audio features are encoded, four attention signals $(w_t, w_{self}, w_{ctx}, w_a)$ jointly weight the video frames, the top-k frames are selected, and a cross-modal transformer fuses them with text and audio to predict misogynistic vs. non-misogynistic content.

quent weight computation:

$$\mathrm{FW}(c, F_v) = \mathrm{softmax}\left(\frac{(W_q\, c)\,(W_k\, F_v)^\top}{\sqrt{d}}\right) \in \mathbb{R}^n. \tag{1}$$

The final temporal weights are derived via multiplicative fusion to extract the top-K ($K = 4$) frames most relevant to the misogynistic context. Note that unlike the original MCA-WF implementation, we rely on standard scaled dot-product attention without a temperature hyperparameter (τ), instead opting for a deterministic top-K extraction:

$$w = \frac{w_{\text{self}} \odot w_{\text{text}} \odot w_{\text{audio}} \odot w_{\text{ctx}}}{\sum(w_{\text{self}} \odot w_{\text{text}} \odot w_{\text{audio}} \odot w_{\text{ctx}})}, \tag{2}$$

Finally, to fuse modalities, we add modality and positional embeddings to the top-K frames (F_k), text (F_t), and audio (F_a): $Z_0 = [F_k + E^{(v)}_{mod}; F_t + E^{(t)}_{mod}; F_a + E^{(a)}_{mod}] + E_{pos}$. The combined sequence passes through $L = 1$ Transformer block, is globally averaged, and classified.

3.4 Training Objective

We optimize a composite loss: $\mathcal{L} = \mathcal{L}_{\text{focal}} + \lambda_{\text{ml}} \left\|F_v - F_v^+\right\|_F^2$, with $\lambda_{\text{ml}} = 0.1$. The focal loss addresses class imbalance dynamically per language. The metric-learning term pulls visual representations toward same-class positives to encourage intra-class compactness.

3.5 Training Regime and Probability Ensembling

To explicitly address severe class imbalances, we implemented a Dynamic Focal Loss that penalizes majority-class bias. All models are trained under a 5-fold cross-validation setup using AdamW (learning rate 10^{-4}, weight decay 10^{-2}) and batch size 8 for up to 100 epochs, with early stopping (patience 10) on validation macro-F1. A fixed random seed (42) controls for initialization variance. During inference, we average the softmax probabilities across the five folds. Ultimately, our final Tamil submissions relied heavily on Gated Early-Fusion variants, while Malayalam submissions utilized a mathematically averaged ensemble of Gated Early-Fusion and Temporal Attention to capture nuanced behaviors.

4 Results and Evaluation

Table 1: Performance on the Malayalam and Tamil test sets.

Lang.	Method	Acc	M-Pre	M-Rec	M-F1
Mal.	Baseline-MCA-WF	0.600	0.639	0.612	0.583
	Baseline-MMHSD	0.675	0.720	0.685	0.665
	TA-Net	0.675	0.720	0.685	0.665
	EF-TA Ensemble	**0.700**	**0.765**	**0.712**	**0.688**
Tam.	Baseline-MCA-WF	0.808	0.810	0.808	0.807
	Baseline-MMHSD	0.808	0.810	0.808	0.807
	EF-Net	0.769	0.776	0.769	0.768
	EF-Net (High-Cap)	**0.808**	**0.810**	**0.808**	**0.807**

On the Malayalam test set, our EF-TA Ensemble significantly outperforms the MMHSD (Céspedes-Sarrias et al., 2025) and MCA-WF (Wang et al., 2023) baselines by 2.3 and 10.5 percentage points (Table 1), confirming the effectiveness of blending early multimodal fusion with temporal attention. Conversely, on the Tamil test set, all top models hit a ceiling Macro-F1 of 0.807. However, class-wise performance reveals these models make distinct precision-recall trade-offs despite identical macro scores.

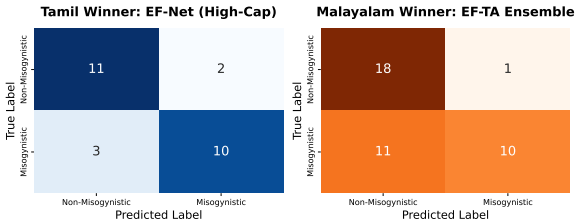


Figure 2: Confusion matrices for the winning architectures on Tamil and Malayalam test sets.

4.1 Cross-Lingual Performance Analysis

A distinct performance difference can be observed in Table 1 between the two languages. The best Macro-F1 obtained for Tamil data is 0.807, whereas the best Macro-F1 for Malayalam data is 0.688, representing a difference of 11.9 percentage points. A similar difference is observed for accuracy: the best Tamil result is 0.808, whereas for Malayalam data it is 0.700. This indicates that the proposed multimodal framework is substantially more effective on the Tamil test set. The observed language-dependent variation may be attributed to differences in dataset characteristics, linguistic properties, modality quality, class distributions, and the effectiveness of the extracted representations for each language. The lower Malayalam performance suggests that further investigation into language-specific representation learning and modality alignment is beneficial.

4.2 Discussion

On the Malayalam set ($n = 40$), the EF-TA Ensemble maximized precision (reducing false positives) at the expense of target-class recall (0.476). This trade-off naturally suppresses noise in constrained datasets, though improving recall remains a priority. On the Tamil set ($n = 26$), the baseline and HighCap variant tied on macro-F1 (0.807) but exhibited diverging operating points (Table 2); the baseline favored recall (84.6% vs. 76.9%) while EF-Net favored precision (83.3% vs. 78.6%).

Table 2: Class-wise Precision (P) and Recall (R) comparing the baselines with the proposed winners.

Language / Model	Class 0 (P / R)	Class 1 (P / R)
Tamil (Constrained)		
- Baseline (MMHSD)	0.833 / 0.769	0.786 / **0.846**
- EF-Net (High-Cap)	0.786 / **0.846**	**0.833** / 0.769
Malayalam (Balanced)		
- Baseline (MMHSD)	0.607 / 0.895	0.833 / 0.476
- **EF-TA Ensemble**	**0.621 / 0.947**	**0.909 / 0.476**

Analyzing local 5-fold validation stability, our Malayalam Temporal Attention network achieved a mean Macro-F1 of 0.917 ± 0.030, demonstrating tighter variance than the MCA-WF baseline (0.887 ± 0.050). The Tamil dataset exhibited predictably high variance ($\approx \pm 0.087$) across all approaches. Ultimately, evaluating on such constrained sets ($n = 26$, $n = 40$) is challenging, as percentage gains resolve to single-video shifts and aggressive regularization forces architectures toward similar decision boundaries. Nevertheless, our framework successfully extracted the incremental advantages necessary to secure Rank 1, indicating strong potential for deployment on larger, unconstrained datasets.

5 Conclusion

We introduced a dual-architecture framework for detecting multimodal misogyny in Dravidian languages. To mitigate catastrophic overfitting on highly constrained datasets, we matched architecture capacity to dataset size. A heavily constrained Gated Early-Fusion Network generalized on the tiny Tamil dataset, while the larger Malayalam dataset supported a deeper Cross-Modal Temporal Attention Network. This empirically driven framework secured Rank 1 across both languages. Because our models were optimized for strict constraints, they favored precision over recall. Future work will expand training data and explore end-to-end joint training to improve recall.

Ethical Considerations

Automated content moderation systems carry inherent risks. Given severe data constraints and the subjectivity of hate speech, our models are susceptible to misclassifications—risking unjust censorship (false positives) and the unchecked spread of harmful content (false negatives). Because our models heavily favor precision, some misogynistic content may evade detection. Furthermore, the limited dataset scale prevents us from guaranteeing the absence of latent socio-cultural biases. We strongly advocate that these models serve strictly as assistive tools for human moderators.

Acknowledgements

Both authors contributed equally to this work.

Code Availability

Our complete code is publicly available at: https://github.com/Projukti-Kavalan/MMLow-2026

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Rui Cao, Roy Ka-Wei Lee, and Tuan-Anh Hoang. 2020. Deephate: Hate speech detection via multi-faceted text representations. In *Proceedings of the 12th ACM Conference on Web Science*, WebSci '20, page 11–20, New York, NY, USA. Association for Computing Machinery.

Berta Céspedes-Sarrias, Carlos Collado-Capell, Pablo Rodenas-Ruiz, Olena Hrynenko, and Andrea Cavallaro. 2025. Mm-hsd: Multi-modal hate speech detection in videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 2546–2555, New York, NY, USA. Association for Computing Machinery.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Navya Jose, Anand Kumar M, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Hariharan R L, John Philip McCrae, and Elizabeth Sherly. 2021. Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 133–145, Kyiv. Association for Computational Linguistics.

Ivan Culjak, David Abram, Tomislav Pribanic, Hrvoje Dzapo, and Mario Cifrek. 2012. A brief introduction to OpenCV. In *2012 Proceedings of the 35th International Convention MIPRO*, pages 1725–1730.

Jyothish Lal G, Premjith B, Bharathi Raja Chakravarthi, Saranya Rajiakodi, Bharathi B, Rajeswari Natarajan, and Ratnavel Rajalakshmi. 2025. Overview of the shared task on multimodal hate speech detection in Dravidian languages: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 114–122, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Nitesh Jindal, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sajeetha Thavareesan, Saranya Rajiakodi, and Bharathi Raja Chakravarthi. 2024. Mistra: Misogyny detection through text–image fusion and representation analysis. *Natural Language Processing Journal*, 7:100073.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. Muril: multilingual representations for indian languages (2021). *arXiv preprint arXiv:2103.10730*.

Gitanjali Kumari, Kirtan Jain, and Asif Ekbal. 2024. M3Hop-CoT: Misogynous meme identification with multimodal multi-hop chain-of-thought. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22105–22138, Miami, Florida, USA. Association for Computational Linguistics.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Shunmuga Priya Muthusamy Chinnan, Ruba Priyadharshini, Rajameenakshi J, Kathiravan P, Rahul Ponnusamy, Bhuvaneswari Sivagnanam, Paul Buitelaar, Bhavanimeena K, Jananayagam V, and Kishore Kumar Ponnusamy. 2025. Findings of the shared task on abusive Tamil and Malayalam text targeting women on social media: DravidianLangTech@NAACL 2025. In *Proceedings of the Fifth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 671–681, Acoma, The Albuquerque Convention Center, Albuquerque, New Mexico. Association for Computational Linguistics.

Simon Rouard, Francisco Massa, and Alexandre Défossez. 2023. Hybrid transformers for music source separation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Yuqing Wang, Zhuang Qi, Xiangxian Li, Jinxing Liu, Xiangxu Meng, and Lei Meng. 2023. Multi-channel attentive weighting of visual frames for multimodal video classification. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

A Appendix: Training Dynamics

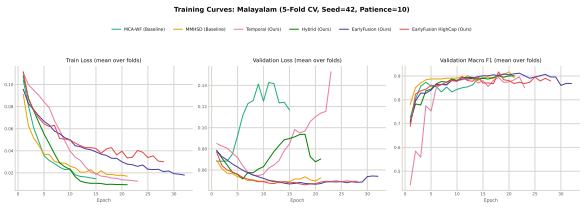


Figure 3: Training/validation loss and validation macro-F1 across epochs for Malayalam (5-fold CV, mean over folds).

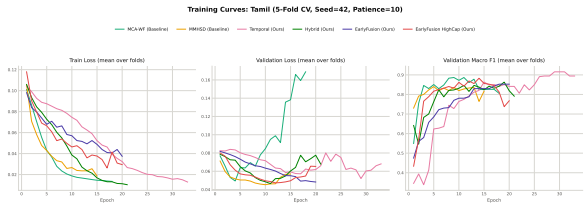


Figure 4: Training/validation loss and validation macro-F1 across epochs for Tamil (5-fold CV, mean over folds).

B Appendix B: Architectural Decision-Making and Ablations

To empirically justify the Empirical Architecture Selection framework described in Section 3, extensive hyperparameter grid searches were conducted on the local validation splits created from the released training dataset. Furthermore, to evaluate the distinct contributions of the data streams, this validation process inherently serves as our modality ablation study. Whereas the existing baseline architectures—Multi-Modal Hate Speech Detection (MMHSD) and Multi-channel Attentive Weighting of Visual Frames (MCA-WF)—utilize restricted modality combinations, our 'Proposed 3-Way' models fully integrate the trimodal features (Audio, Visual, and Text). While a strict unimodal Text-Only baseline was not evaluated due to the shared task's submission limits, the performance gap between the restricted-modality baselines and our trimodal architectures demonstrates the added value of the visual and audio streams, as detailed in Table 3.

B.1 Text Encoder Selection

We systematically evaluated two state-of-the-art multilingual text encoders: E5-Large and MuRiL-Large (optimized for Indian languages). Across both datasets and irrespective of whether a Pooled (Early Fusion) or Temporal (Cross-Modal Attention) architecture was utilized, the MuRiL-Large embeddings consistently outperformed E5-Large. Notably, on the Tamil dataset, the MuRiL-Large Temporal setup with our proposed 3-way routing achieved a validation macro-F1 of 1.000 (zero false positives/negatives on the local fold), whereas E5-Large capped at 0.956. For Malayalam, the MuRiL-Large Pooled configuration achieved a local validation macro-F1 of 0.917, significantly surpassing E5-Large's 0.852. Consequently, MuRiL-Large was selected as the definitive text extractor for all final models. Note that the perfect validation score of 1.000 for the Temporal model on Tamil is a small-sample validation artifact representing catastrophic overfitting to a single 24-video fold, which justified our shift to the constrained EF-Net.

B.2 Hyperparameter Grid Search and Regularization

The severe disparity in dataset sizes required careful tuning of our regularization strategies to prevent catastrophic memorization, particularly on the Tamil dataset.

Standard Architectures (EF-Net & TA-Net): For our core architectures—the Gated Early-Fusion Network (EF-Net) and the Cross-Modal Temporal Attention Network (TA-Net)—an extensive grid search revealed that aggressive compression and regularization were universally necessary across both languages to maintain out-of-sample robustness. The optimal configuration for both Tamil and Malayalam utilized a unified latent dimension of $DIM = 64$, an extreme dropout rate of 0.7, and a high weight decay of 0.01. This massive informational bottleneck proved highly effective at suppressing dataset-specific noise.

High-Capacity Variant: To explicitly test if the Tamil dataset could support deeper representations without overfitting, we evaluated the `EF-Net (High-Cap)` variant. This model was configured with a much wider latent dimension of $DIM = 512$ and a more forgiving dropout of 0.4. While this variant successfully reached parity with the baselines on the Tamil test set, the standard heavily-regularized architectures proved more stable across both languages on average, as shown by the variance metrics in Table 4.

C Appendix C: Final Competition Submissions

During the official final phase of the competition, we submitted only our top 3 models per language based on their validation performance. However, to provide a comprehensive scientific analysis for this paper, Table 5 reports the test set results for an expanded suite of 9 runs per language, representing a systematic progression from established baselines to our specialized multimodal ensembles. This clarifies that our Rank 1 placement was achieved within a strict 3-submission limit, and the expanded 9-run post-evaluation confirms that the EF-Net (High-Cap) variant and the EF-TA Ensemble provide the most robust generalization.

Table 3: Comprehensive modality and architecture ablation results evaluated on the local validation split. This demonstrates modality ablation: while baselines rely on limited modalities, the 'Proposed 3-Way' achieves performance gains via full trimodal integration (Audio, Visual, Text). Note that the 1.000 score on Tamil is a localized small-sample validation artifact.

Language	Text Encoder	Input Type	Model Architecture	Val Accuracy	Val Macro-F1
Tamil	E5-Large	Pooled	MMHSD	0.833	0.829
Tamil	E5-Large	Pooled	MCA-WF	0.958	0.955
Tamil	E5-Large	Pooled	Proposed 3-Way	0.917	0.911
Tamil	E5-Large	Temporal	MMHSD	0.875	0.863
Tamil	E5-Large	Temporal	MCA-WF	0.875	0.873
Tamil	E5-Large	Temporal	Proposed 3-Way	0.958	0.956
Tamil	MuRiL-Large	Pooled	MMHSD	0.917	0.914
Tamil	MuRiL-Large	Pooled	MCA-WF	0.958	0.956
Tamil	MuRiL-Large	Pooled	Proposed 3-Way	0.917	0.914
Tamil	MuRiL-Large	Temporal	MMHSD	0.875	0.863
Tamil	MuRiL-Large	Temporal	MCA-WF	0.875	0.869
Tamil	MuRiL-Large	Temporal	Proposed 3-Way	1.000	1.000
Malayalam	E5-Large	Pooled	MMHSD	0.853	0.853
Malayalam	E5-Large	Pooled	MCA-WF	0.853	0.852
Malayalam	E5-Large	Pooled	Proposed 3-Way	0.853	0.852
Malayalam	E5-Large	Temporal	MMHSD	0.836	0.836
Malayalam	E5-Large	Temporal	MCA-WF	0.787	0.787
Malayalam	E5-Large	Temporal	Proposed 3-Way	0.869	0.869
Malayalam	MuRiL-Large	Pooled	MMHSD	0.902	0.901
Malayalam	MuRiL-Large	Pooled	MCA-WF	0.918	0.918
Malayalam	MuRiL-Large	Pooled	Proposed 3-Way	0.918	0.918
Malayalam	MuRiL-Large	Temporal	MMHSD	0.836	0.836
Malayalam	MuRiL-Large	Temporal	MCA-WF	0.869	0.869
Malayalam	MuRiL-Large	Temporal	Proposed 3-Way	0.869	0.869

Table 4: Mean local validation Macro-F1 scores and standard deviations across 5 folds for our proposed architectures and baselines (utilizing MuRiL-Large embeddings). The variance highlights the stability of our models on Malayalam and the extreme volatility inherent to the tiny Tamil dataset.

Language	Model	Macro-F1 (Mean $\pm$ SD)
Malayalam	Baseline: MCA-WF	0.887 ± 0.050
Malayalam	Baseline: MMHSD	0.930 ± 0.026
Malayalam	Proposed: Early Fusion	0.917 ± 0.060
Malayalam	Proposed: Temporal (TA)	0.917 ± 0.030
Tamil	Baseline: MCA-WF	0.901 ± 0.085
Tamil	Baseline: MMHSD	0.868 ± 0.086
Tamil	Proposed: Temporal (TA)	0.900 ± 0.085
Tamil	Proposed: EF-Net (High-Cap)	0.887 ± 0.087

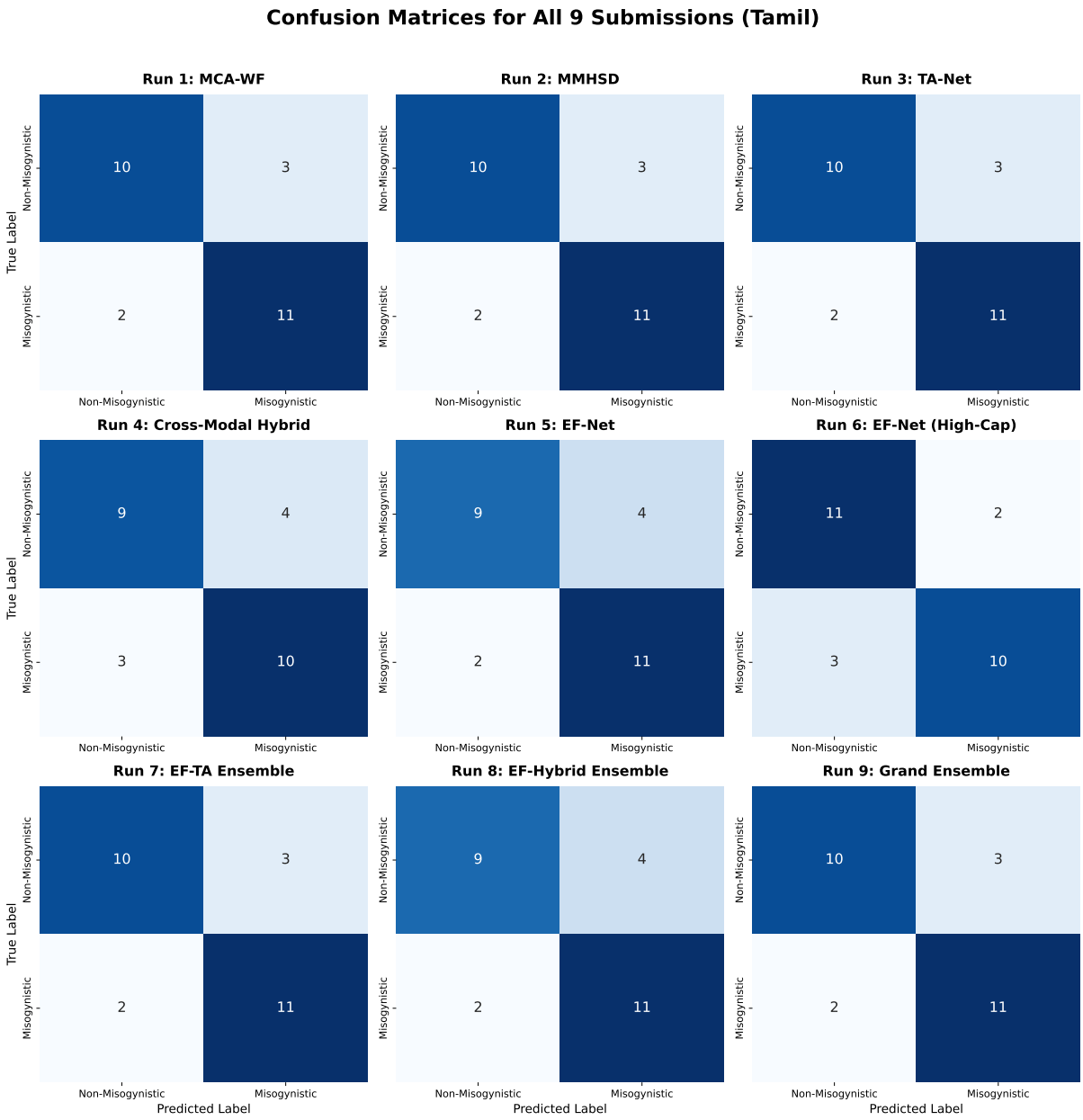


Figure 5: Detailed confusion matrices for all 9 models submitted for the Tamil dataset, demonstrating the evolution of predictive behavior.

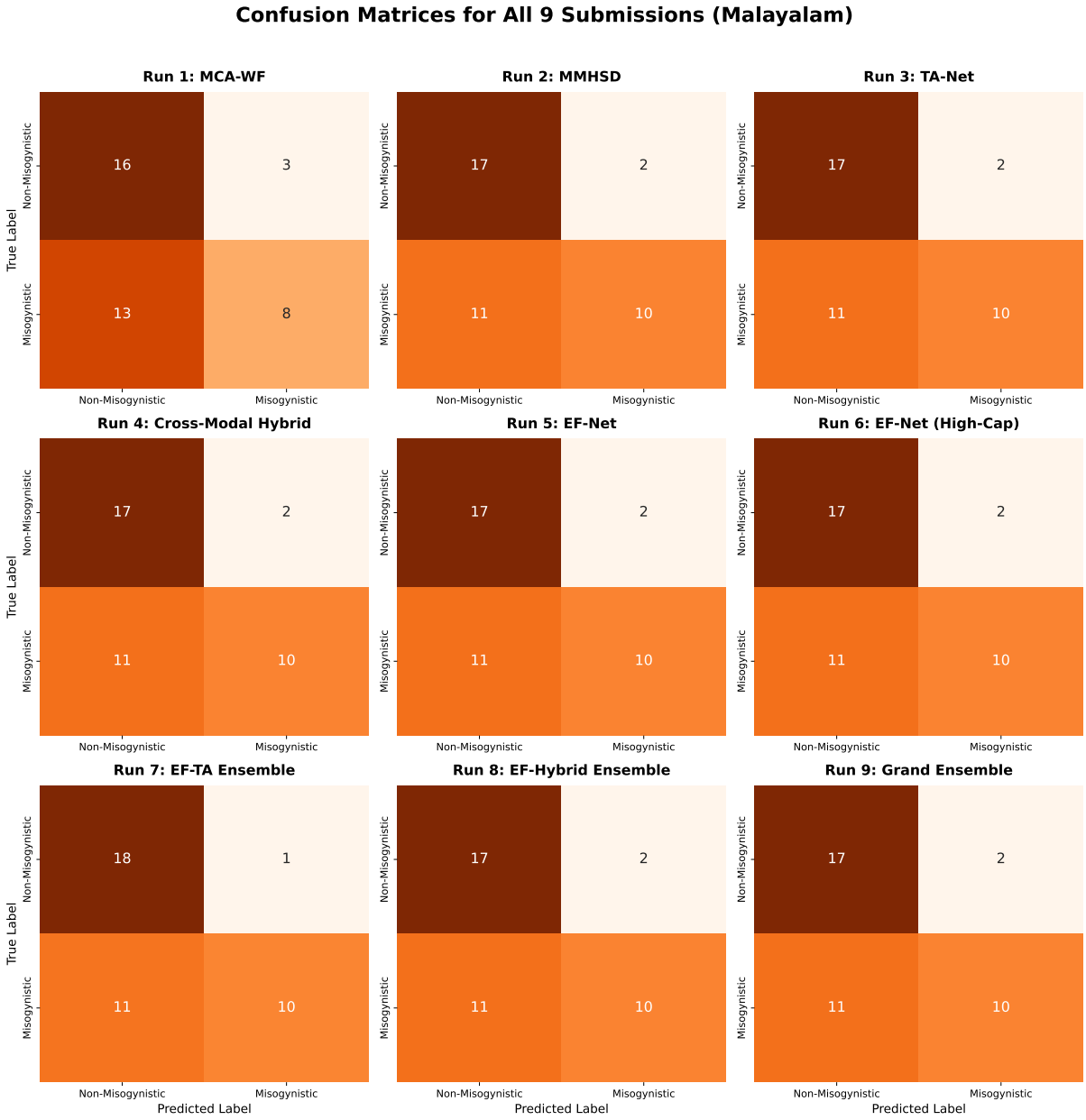


Figure 6: Detailed confusion matrices for all 9 models submitted for the Malayalam dataset, demonstrating the evolution of predictive behavior.

Table 5: Results of all 9 experimental runs submitted to the final evaluation phase for both datasets.

Language	**Model Architecture**	**Accuracy**	**Macro-Precision**	**Macro-Recall**	**Macro-F1**
Tamil	MCA-WF	0.808	0.810	0.808	0.807
Tamil	MMHSD	0.808	0.810	0.808	0.807
Tamil	TA-Net	0.808	0.810	0.808	0.807
Tamil	Cross-Modal Hybrid	0.731	0.732	0.731	0.730
Tamil	EF-Net	0.769	0.776	0.769	0.768
Tamil	EF-Net (High-Cap)	0.808	0.810	0.808	0.807
Tamil	EF-TA Ensemble	0.808	0.810	0.808	0.807
Tamil	EF-Hybrid Ensemble	0.769	0.776	0.769	0.768
Tamil	Grand Ensemble	0.808	0.810	0.808	0.807
Malayalam	MCA-WF	0.600	0.639	0.612	0.583
Malayalam	MMHSD	0.675	0.720	0.685	0.665
Malayalam	TA-Net	0.675	0.720	0.685	0.665
Malayalam	Cross-Modal Hybrid	0.675	0.720	0.685	0.665
Malayalam	EF-Net	0.675	0.720	0.685	0.665
Malayalam	EF-Net (High-Cap)	0.675	0.720	0.685	0.665
Malayalam	EF-TA Ensemble	0.700	0.765	0.712	0.688
Malayalam	EF-Hybrid Ensemble	0.675	0.720	0.685	0.665
Malayalam	Grand Ensemble	0.675	0.720	0.685	0.665