RMS@DravidianLangTech 2026: Multimodal Gated Fusion for Hierarchical Tamil Political Meme Classification

Md. Ajwad Hossain
Department Of Electronics & Telecommunication Engineering
Chittagong University Of Engineering & Technology
md.ajwadhossain@gmail.com

Abstract

Internet memes have become a dominant and highly accessible medium for political discourse on social media. However, their multimodal nature—combining culturally specific visual symbols with code-mixed text—presents a significant challenge for automated content analysis, particularly in low-resource languages. In this study, we describe the system submitted by team RMS for the Multi-Level Political Meme Classification shared task at DravidianLangTech @ ACL 2026, focusing exclusively on the Tamil language track. We propose a robust late-fusion multimodal architecture that leverages a pre-trained ResNet-50 network for visual feature extraction and a Transformer-based model (MuRIL) for processing code-mixed Tamil text. The modalities are aligned using bidirectional cross-modal attention and combined using a Gated Multimodal Unit, allowing the model to dynamically weight the importance of visual versus textual cues. Our system ranked 11th on the official leaderboard with a macro-averaged F1-score of 0.7382. Through detailed error analysis, we demonstrate that while our gated fusion approach excels at identifying explicit trolling stances, it struggles with complex target resolution when visual and textual cues contradict.

1 Introduction

The rapid proliferation of political memes on social media has transformed how political narratives are shaped, consumed, and heavily debated. Unlike traditional text-based posts, memes rely on a complex, often sarcastic interplay between visual elements—such as facial expressions, background contexts, and political symbols—and short textual overlays (Kiela et al., 2020). This multimodal combination allows memes to convey rich political stances seamlessly, but it presents a formidable challenge for automated content moderation, particularly in low-resource and highly code-mixed languages like Tamil (Suryawanshi et al., 2020).

To address this gap, the DravidianLangTech @ ACL 2026 workshop introduced the Multi-Level Political Meme Classification shared task (Rajiakodi et al., 2026). The task requires systems to perform a hierarchical classification: first determining the overall stance of the meme (Support vs. Troll/Oppose), and subsequently identifying the specific target of that stance (Individual Person vs. Political Party).

In this study, team RMS presents a robust, multimodal framework designed specifically for the Tamil track of this shared task. While recent advancements in multimodal learning have explored complex cross-attention mechanisms, these models are computationally expensive and highly prone to overfitting on small, noisy datasets. Therefore, we investigate the efficacy of a stable late-fusion approach. Our architecture independently extracts features from the meme's image using a deep residual network (ResNet-50) and the embedded Tamil text using a multilingual Transformer (MuRIL). These representations are subsequently integrated using a Gated Multimodal Fusion layer, which dynamically adjusts the contribution of each modality based on its semantic relevance.

Our empirical results demonstrate that this gated architecture provides a highly competitive baseline for the hierarchical classification task, achieving an overall macro-averaged F1-score of 0.7382 and ranking 11th on the official leaderboard. We make our code publicly available to support future research in low-resource multimodal content analysis. [1]

2 Related Work

Multimodal Meme Analysis: The automated detection of trolls, hate speech, and sentiment in internet memes has gained significant traction. Early approaches primarily relied on unimodal text analysis,

[1] Code is available on GitHub.

Proceedings of the Sixth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 341–347

applying standard NLP techniques to extracted Optical Character Recognition (OCR) text. However, Kiela et al. (2020) demonstrated that memes are often "benign confounders," where the text and image are harmless independently but toxic when combined, necessitating true multimodal architectures. In the context of Dravidian languages, Suryawanshi et al. (2020) pioneered the creation of Tamil meme datasets, highlighting the unique challenges of parsing code-mixed Tanglish text alongside culturally specific visual humor.

Code-Mixed Text and Visual Encoding: Processing Tanglish text requires models capable of handling transliteration and code-switching without relying on strict syntactic rules. Khanuja et al. (2021) introduced MuRIL, a BERT-based representation explicitly pre-trained on Indian languages, which has become the standard for such tasks. For visual encoding, deep convolutional networks like ResNet (He et al., 2016) remain the backbone for extracting high-level features from noisy meme templates.

Multimodal Fusion Strategies: The crux of meme classification lies in how the modalities are fused. Simple concatenation (early fusion) often suffers from modality dominance, where a strong visual signal overshadows textual nuances. While recent transformer-based cross-attention models align modalities effectively, they require massive datasets to train. As an alternative, Arevalo et al. (2017) proposed Gated Multimodal Units (GMU), which learn to selectively weight visual and textual features. Our work adapts this gating philosophy, utilizing it to stabilize training on a highly imbalanced, low-resource political meme dataset.

3 Task and Dataset Description

The Multi-Level Political Meme Classification shared task (Rajiakodi et al., 2026) provides a dataset of political memes in Dravidian languages. Our team participated exclusively in the Tamil language track. The task is structured hierarchically into two distinct levels of classification:

Level 1 (Stance Detection): This primary task requires classifying the overall sentiment or stance of the multimodal meme. The system must predict whether the meme expresses a 'Support' stance or a 'Troll/Oppose' stance towards its subject.

Level 2 (Target Identification): For memes identified in Level 1, this secondary task requires fine-grained classification to determine the specific target of the stance. The target classes include broad categories such as 'Individual Person' and 'Political Party'.

The dataset presents unique challenges, as the memes feature culturally specific Tamil political imagery combined with code-mixed Tanglish (Tamil and English) text, often utilizing heavy sarcasm where the visual and textual modalities may present contradictory sentiments. A critical challenge of this dataset is the severe class imbalance present in the training distribution. In Level 1, the data is heavily skewed toward the *Troll/Oppose* class (691 instances) compared to the *Support/Praise* class (112 instances). This imbalance cascades into Level 2, where *Troll/Oppose Against Person* (547 instances) heavily dominates minority classes such as *Support for party* (24 instances), requiring the model to learn robust representations from very few positive examples.

4 Methodology

Our proposed system employs a sophisticated late-fusion multimodal architecture to hierarchically classify political memes. To address the semantic gap between visual and textual modalities, we extract deep representations independently and align them using a Bidirectional Cross-Modal Attention mechanism. The aligned features are then dynamically weighted using a Gated Multimodal Unit before being passed to hierarchical classification heads.

4.1 Text Extraction and Preprocessing

Since internet memes embed text directly within the image payload, robust Optical Character Recognition (OCR) is a prerequisite for multimodal learning. To extract the code-mixed Tamil and English text from the noisy meme backgrounds, we utilized the `ocr_tamil` library, which is specifically optimized for Dravidian scripts (Baek et al., 2019; Bautista and Atienza, 2022; D, 2024).

The raw OCR output frequently contains internet artifacts and structural noise that degrade the performance of transformer-based language models. To mitigate this, we implemented a comprehensive text-cleaning pipeline. Our preprocessing sequence systematically removes URLs, email addresses, and non-alphanumeric special characters. Additionally, we normalized excessive whitespaces, converted all text to lowercase, and truncated consecutive duplicate characters—a common fea-

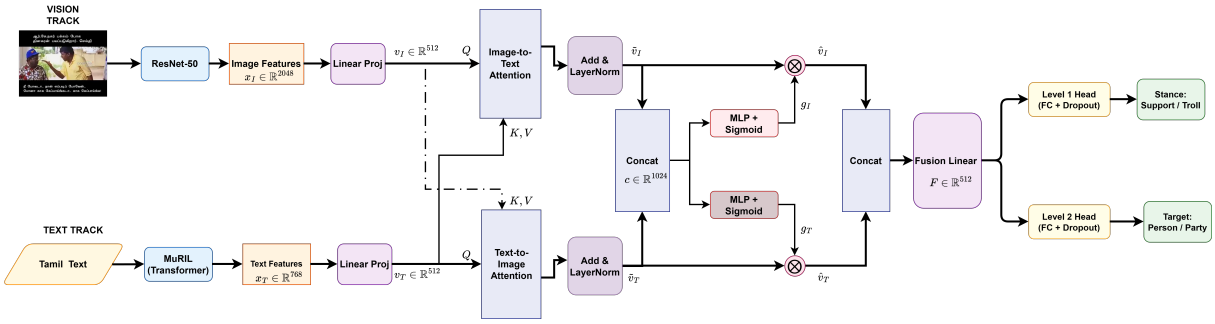

Figure 1: The proposed Gated Cross-Modal architecture utilized by Team RMS. The model aligns visual and textual features via cross-attention and dynamically fuses them using modality-specific gating mechanisms.

ture in memes used for comedic emphasis (e.g., "loool")—to a maximum of two repetitions. In instances where the OCR engine failed to detect any text, a fallback generic token (`meme_image`) was supplied to the textual encoder to ensure dimensional stability during batch processing.

4.2 Modality Encoders

Visual Encoder: To capture the visual semantics of the memes, including facial expressions and political symbols, we utilize a pre-trained ResNet-50 network. We remove the final classification head and extract the flattened output from the global average pooling layer. For a given input image I, the raw visual feature vector is:

$$x_I = \text{ResNet50}(I) \quad \text{where} \quad x_I \in \mathbb{R}^{2048} \tag{1}$$

Textual Encoder: The textual components of the memes are highly code-mixed Tamil (Tanglish) and English. We employ MuRIL, a Transformer model specifically pre-trained on Indian text corpora. For a given caption T, we tokenize the sequence and extract the pooled `[CLS]` token embedding:

$$x_T = \text{MuRIL}(T) \quad \text{where} \quad x_T \in \mathbb{R}^{768} \tag{2}$$

4.3 Bidirectional Cross-Modal Attention

Before fusing the modalities, we align them into a shared semantic space to allow the visual and textual features to inform one another. The raw modality vectors are linearly projected into a shared dimension $d = 512$:

$$v_I = W_{pI} x_I + b_{pI}, \quad v_T = W_{pT} x_T + b_{pT} \tag{3}$$

We employ residual connections and Layer Normalization (LN). Let $\text{MHA}(Q, K, V)$ denote the Multi-Head Attention operation. The cross-attended features are computed as:

$$\tilde{v}_I = \text{LN}(v_I + \text{MHA}(v_I, v_T, v_T)) \tag{4}$$

$$\tilde{v}_T = \text{LN}(v_T + \text{MHA}(v_T, v_I, v_I)) \tag{5}$$

Here, $\tilde{v}_I$ and $\tilde{v}_T \in \mathbb{R}^{512}$ represent the cross-attended image and text features.

4.4 Gated Multimodal Fusion

To dynamically control the contribution of each modality, we introduce a Gated Fusion module. We concatenate the attended features to form a joint representation $c \in \mathbb{R}^{1024}$:

$$c = [\tilde{v}_I; \tilde{v}_T] \tag{6}$$

This representation is passed through two parallel Multi-Layer Perceptrons (MLPs) to generate modality-specific scalar gates. Each MLP uses a Sigmoid activation (σ) to constrain the output to $(0, 1)$:

$$g_I = \sigma(W_{gI2}(\text{ReLU}(W_{gI1} c + b_{gI1})) + b_{gI2}) \tag{7}$$

$$g_T = \sigma(W_{gT2}(\text{ReLU}(W_{gT1} c + b_{gT1})) + b_{gT2}) \tag{8}$$

The attended features are then scaled by their respective learned gates:

$$\hat{v}_I = \tilde{v}_I \cdot g_I, \quad \hat{v}_T = \tilde{v}_T \cdot g_T \tag{9}$$

Finally, the gated features are concatenated and linearly projected back to $d = 512$ to form the unified multimodal embedding F:

$$F = W_F[\hat{v}_I; \hat{v}_T] + b_F \tag{10}$$

4.5 Hierarchical Classification

The unified representation F is fed into two parallel classification heads. Each head consists of a fully connected layer, a ReLU activation, Dropout ($p = 0.3$), and a final linear projection.

For the Level 1 task (Stance Detection), the probability distribution over the classes is computed as:

$$h_{L1} = \text{Dropout}(\text{ReLU}(W_{1a}F)) \quad (11)$$

$$\hat{p}_{L1} = \text{softmax}(W_{1b}h_{L1} + b_{1b}) \quad (12)$$

A similar formulation is used to yield $\hat{p}_{L2}$ for the Level 2 task. The network is optimized jointly using an unweighted sum of categorical cross-entropy losses:

$$\mathcal{L}_{total} = \mathcal{L}_{CE}(\hat{y}_{L1}, y_{L1}) + \mathcal{L}_{CE}(\hat{y}_{L2}, y_{L2}) \quad (13)$$

5 Results and Error Analysis

5.1 Experimental Results

Implementation Details: All architectural variants were trained for 15 epochs using a learning rate of 4×10^{-5}. To ensure strict reproducibility across all internal validations and final test-set evaluations, the random seed was universally fixed to 42.

Table 1 presents the official test set performance of our Gated Multimodal Fusion system. The model achieved an F1-score of 0.8608 on the Level 1 stance detection task, demonstrating a strong capability to distinguish between standard Support and Troll memes. Performance on the more granular Level 2 target identification task yielded an F1-score of 0.6156, reflecting the inherent difficulty of resolving specific political targets in highly code-mixed, low-resource data. Overall, the system achieved a macro-averaged F1 of 0.7382, securing the 11th position among the participating systems.

While we conducted preliminary internal evaluations and hyperparameter tuning using a standard train-validation split, we found that unimodal and un-gated architectures were prone to rapid overfitting. As detailed in our comprehensive ablation study (see Appendix A), the decision to employ a gated cross-modal approach provided the necessary regularization to maintain stable validation trajectories. Furthermore, visual evidence of model capacity from these ablation curves strictly motivated our final submission strategy: retraining the selected architecture on the entirety of the provided dataset to maximize exposure to minority class features before generating predictions for the official test set.

Task	Precision	Recall	F1-Score
L1 (Stance)	0.8554	0.8756	0.8608
L2 (Target)	0.6151	0.6318	0.6156
Avg F1	**0.7382 (Rank: 11th)**		

Table 1: Official test set results for the RMS model.

5.2 Error Analysis

Class Imbalance: The most prominent source of error stems from the severe skew in the training data. The model demonstrated excellent recall for the dominant *Troll/Oppose* class (which constitutes roughly 86% of the training data) but struggled significantly with the minority *Support* classes. When cross-modal signals were ambiguous, the model heavily favored the majority prior, frequently defaulting to a Troll prediction.

Target Confusion (Person vs. Party): In Level 2, the model frequently misclassified 'Troll against Party' as 'Troll against Individual Person'. Memes criticizing a political party almost always feature a recognizable photograph of that party's leader. Because our visual encoder is adept at facial feature extraction, the presence of a face strongly triggered the visual pathway to output representations aligned with the 'Individual Person' class. The textual encoder occasionally failed to extract the specific party-level nuances from the Tanglish text to override this strong visual signal.

6 Conclusion

In this study, we described the system submitted by team RMS for the Multi-Level Political Meme Classification shared task. We designed a late-fusion multimodal architecture that aligns deep visual and textual representations using Bidirectional Cross-Modal Attention, followed by a Gated Multimodal Unit. Our approach achieved competitive results, ranking 11th on the official Tamil track leaderboard. Our analysis revealed a distinct bias in the visual pathway, where the presence of a politician's face often caused the model to misclassify party-directed trolling. Future work should focus on explicit entity-grounding techniques that allow textual cues to override dominating visual signals.

Limitations

While our gated fusion architecture demonstrates competitive performance, several limitations exist. First, the model is trained and evaluated exclusively on the provided Tamil dataset; its generalizability to the Malayalam track or other cross-

lingual zero-shot scenarios remains untested. Second, as highlighted in our error analysis, the visual encoder exhibits a strong prior toward facial features, which can overpower textual cues in complex code-mixed scenarios. Addressing this requires more sophisticated visual-textual grounding mechanisms. Finally, the severe class imbalance restricts the model's reliability on minority classes (e.g., 'Support for party'), indicating that future work must explore advanced re-weighting strategies or synthetic data generation for low-resource Tanglish text.

Finally, while our gated late-fusion approach establishes a computationally efficient baseline, future iterations would significantly benefit from exploring recent advancements in Multimodal Large Language Models (MLLMs). Foundation models such as CLIP (Radford et al., 2021) and LLaVA (Liu et al., 2023) offer promising avenues for transfer learning, which could improve generalizability across low-resource Dravidian languages. Leveraging these large-scale vision-language models could potentially resolve the modality dominance issues we observed by providing deeper semantic grounding when visual and textual cues contradict.

Ethical considerations

The automated classification of political memes carries significant ethical implications. While systems like ours are designed to assist in content moderation and prevent the spread of malicious trolling, they must be deployed with caution to avoid silencing legitimate political dissent or satire. Furthermore, we acknowledge that the dataset contains offensive content; all data was processed strictly for research purposes following the shared task guidelines.

Finally, to prevent misuse, the approach should be viewed as a human-in-the-loop assistive tool rather than an autonomous moderation system.

References

John Arevalo, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González. 2017. Gated multimodal units for information fusion. *Preprint*, arXiv:1702.01992.

Youngmin Baek, Bado Lee, Dongyoon Han, Sangdoo Yun, and Hwalsuk Lee. 2019. Character region awareness for text detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9365–9374.

Darwin Bautista and Rowel Atienza. 2022. Scene text recognition with permuted autoregressive sequence models. In *European Conference on Computer Vision*, pages 178–196, Cham. Springer Nature Switzerland.

Gnana Prasath D. 2024. Tamil ocr.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: detecting hate speech in multimodal memes. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Saranya Rajiakodi, Shunmuga Priya Muthusamy Chinnan, Premjith B, Subalalitha CN, Rahul Ponnusamy, Anshid K A, Bhuvaneshwari Sivagnanam, Jananayagan V, Bharathi Raja Chakravarthi, Ragavan N, and Santhini P. 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.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

A Ablation Study and Training Dynamics

To isolate the contributions of the visual and textual modalities, we conducted a comprehensive ablation study utilizing a 25% internal validation split. As shown in Table 2, we evaluated four architectural variations, recording both accuracy and macro F1-scores at the final epoch.

Architecture	Level 1		Level 2	
	Acc	F1	Acc	F1
Image-Only	0.9104	0.8997	0.6965	0.6511
Attention-Only	0.9104	0.8973	0.6915	0.6261
Gated Fusion	**0.9204**	**0.9109**	**0.6915**	**0.6470**
Cross-Modal (Ours)	0.9104	0.8973	0.6816	0.6421

Table 2: Final epoch validation metrics for architectural variants.

Due to the limited size of the validation split and the severe visual bias inherent in the meme dataset, performance metrics exhibited high variance. As illustrated in Figure 2, plotting the training versus validation accuracy on a fixed scale reveals severe overfitting trajectories across all models. Training accuracy rapidly converges to 1.0 by Epoch 5, while validation accuracy plateaus early.

This pronounced train-validation gap strongly indicated that our models possessed excess capacity for the small training set. Consequently, while the Gated Cross-Modal architecture was selected for its theoretical capacity to resolve complex visual-textual contradictions, this visual evidence strictly necessitated our final competition strategy: retraining the selected model on the entirety of the combined dataset before inference to maximize exposure to the highly skewed minority classes.

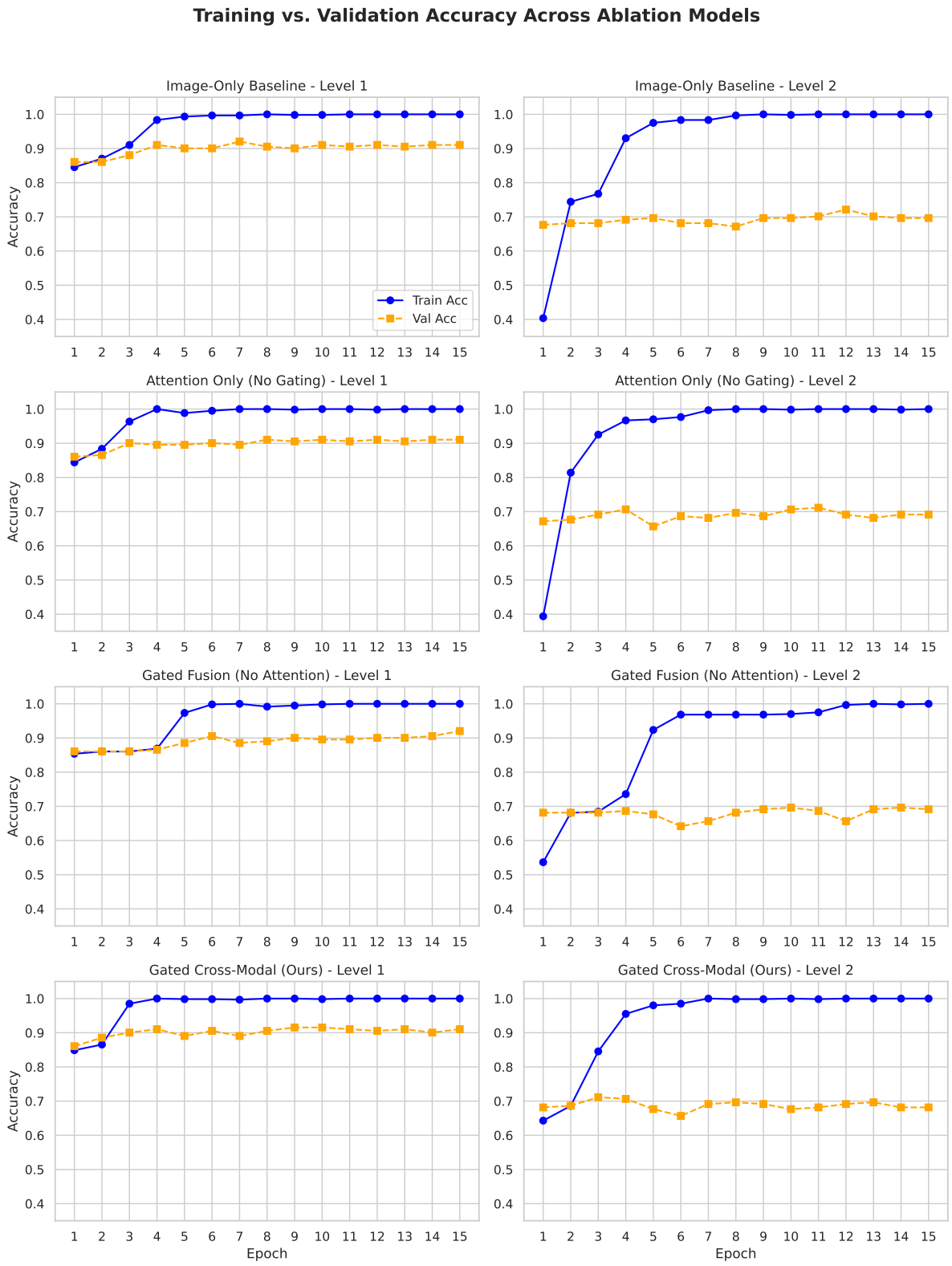

Figure 2: Training versus Validation Accuracy across 15 epochs for all four architectural variants (Level 1 left, Level 2 right). The curves demonstrate the aggressive overfitting dynamic, motivating our final submission strategy of retraining the selected architecture on the available training data.