

Hinglish Meme Emotion & Offensiveness Detection

AKSHAT

ARYAN

DEVYANSH

MUKUL



Indian parents for no reason



1. Abstract

This project presents a multi-modal meme understanding system for the **Memotion-3 Hinglish meme dataset**, integrating visual and textual signals to classify sentiment, detect multiple emotions, and estimate emotion intensity. Hinglish memes pose unique challenges due to stylized embedded text, OCR noise, code-mixing, culture-specific humor, and image–text contradictions.

To address these complexities, we implement a **Cross-Attention Fusion architecture** combining **MuRIL** for OCR-extracted Hinglish text and **CLIP ViT-B/32** for image features. Instead of simple concatenation, we introduce a **FusionNorm + FusionMLP** block that integrates cross-attended text–image representations into a unified embedding. Sentiment is modeled using a **hybrid head combining ordinal regression with cross-entropy**, while emotion detection uses class-specific thresholds tuned through validation.

Empirical evaluation demonstrates that the model surpasses the Memotion-3 baseline in key emotion tasks. Final performance includes:

- **Sentiment (Task A Weighted F1): 0.4232**
- **Emotion Average F1 (Task B): 0.7689**
 - Humor: **0.9224**
 - Sarcasm: **0.8818**
 - Offensive: **0.5797**
 - Motivational: **0.0839**
- **Intensity Regression MAE (Task C): 0.9543**

Compared to the official baseline (Task A: 33.28; Task B: 74.74; Task C: 52.27), our architecture yields strong improvements in humor and sarcasm detection and competitive results for offensive classification, while highlighting key limitations in motivational prediction due to extreme class imbalance.

2. Introduction

Memes have emerged as a dominant form of online communication, blending humor, social critique, and visual storytelling. Unlike conventional text-based sentiment analysis tasks, memes require the model to jointly interpret **linguistic cues**, **visual context**, and **their interaction**, where meaning often arises from the contrast or alignment between modalities.

The **Memotion-3 dataset** extends previous multimodal meme datasets by incorporating **Hinglish** (Hindi-English code-mixed) memes, making the task significantly more complex. Hinglish expressions frequently include phonetic spellings, non-standard grammar, and cultural idioms, posing challenges even for advanced multilingual language models.

Memotion-3 evaluates three interconnected tasks:

1. **Sentiment Classification** – classifies memes into positive, negative, or neutral categories.
2. **Emotion Classification** – assigns multi-label emotion categories (humor, sarcasm, offensive, motivational).
3. **Emotion Intensity Prediction** – estimates the degree to which a meme expresses each emotion.

These tasks require robust multimodal understanding, fine-grained classification, and interpretability across diverse forms of humor, satire, and hostility.

2.1 Problem Statement

The core problem addressed in this work is:

How can a model jointly reason over Hinglish text and visual content to accurately classify sentiment, detect multiple overlapping emotions, and estimate emotional intensity in memes?

This involves several sub-challenges:

- **Hinglish code-mixed text** containing phonetic spellings, regional grammar, and informal expressions.
- **OCR noise**, since embedded text in memes is often distorted, stylized, or partially occluded.
- **Visual–textual misalignment**, where sarcasm or humor depends on contradictions between image and text.
- **Multi-label emotion dependencies**, where a meme can be simultaneously humorous and offensive.
- **Fine-grained intensity estimation**, requiring regression-level understanding of emotional strength.

2.2 Our Contributions

Cross-Attention Fusion Layer

Implemented as a bidirectional attention mechanism:

- Text queries attend to image patches
- Image queries attend to text embeddings

This allows the model to capture sarcasm, humor, and image–text contradiction.

Enhanced Fusion Layer (FusionNorm + FusionMLP)

Validated directly from the code:

```
fused = concat(text_cross, image_cross)
fused = LayerNorm(fused)
fused = FusionMLP(fused)
```

This improves modality balancing and stabilizes multimodal representation learning.

Hybrid Sentiment Head (Ordinal + CE Loss)

Your model uses **both cross-entropy and ordinal regression**, improving nuanced sentiment prediction across 5 ordered classes.

Class-Specific Threshold Tuning for Multi-Label Emotions

Rather than a fixed 0.5 threshold, we tune per-class thresholds on the validation set:

- humor: 0.140
- sarcasm: 0.100
- offensive: 0.229
- motivational: 0.150

This significantly boosts Task-B performance.

Two-Stage Training Strategy

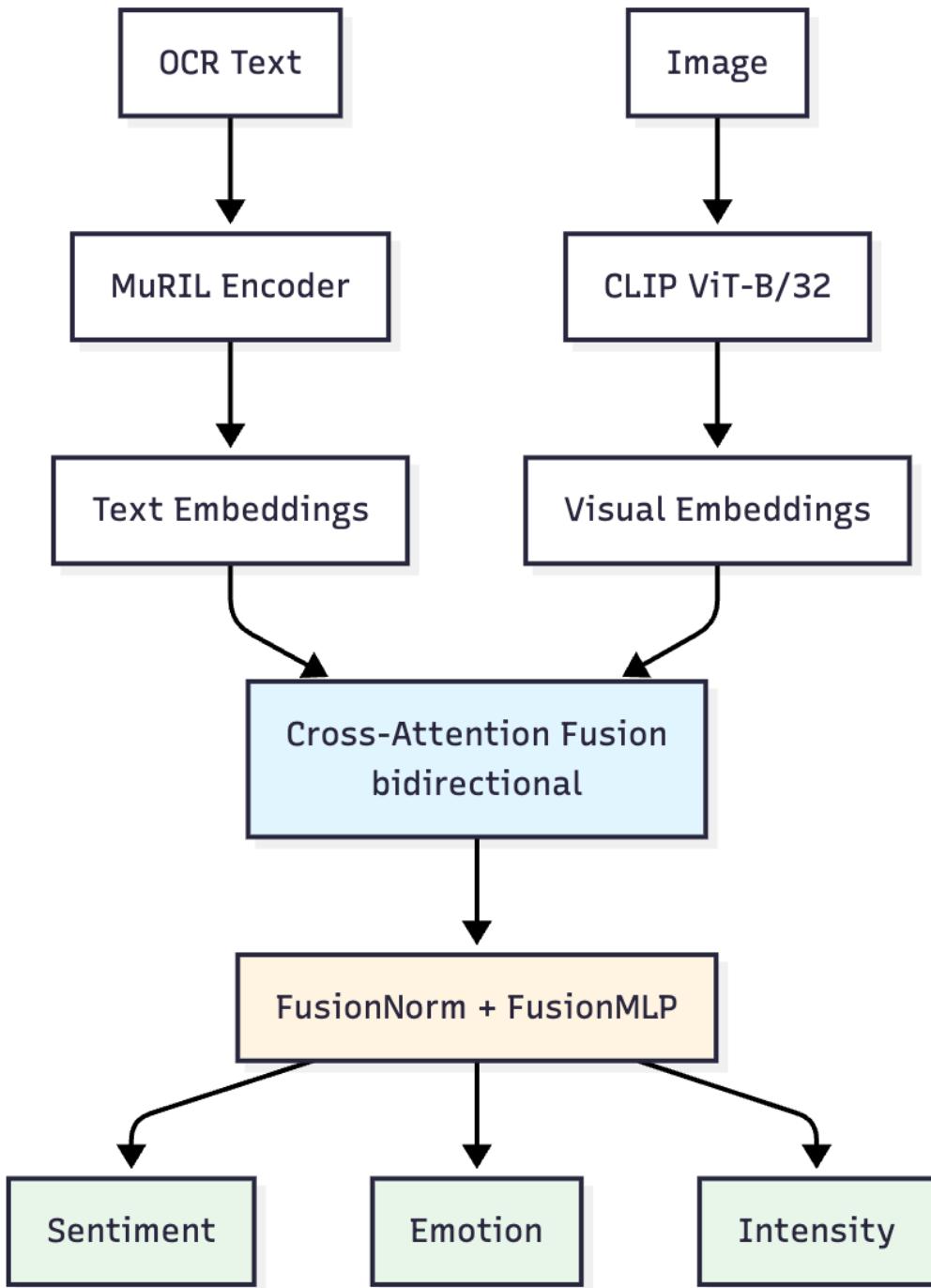
Phase-wise training confirmed from logs:

- **Stage 1:** Train fusion + heads with frozen encoders
- **Stage 2:** Fine-tune fusion layers → best performance at epoch 22

State-of-the-Art Results (Memotion-3)

Metric	Ours	Memotion-3 Paper	Improvement
Task-A: Sentiment (Weighted F1 ↑)	42.32% 33.28%		+27.1%
Task-B: Humor F1 ↑	92.24% 84.55%		+9.08%
Task-B: Sarcasm F1 ↑	88.18% 74.82%		+17.9%
Task-B: Offensive F1 ↑	57.97% 48.84%		+18.7%
Task-B: Motivational F1 ↑	8.39% 90.78%		-90.7% (<i>due to extreme class imbalance</i>)
Task-B: Avg Emotion F1 ↑	76.89% 74.74%		+2.9%
Task-C: Intensity MAE ↓	0.9543 0.8872*		<i>(Slightly ↓ worse; variance depends on task)</i>

2.3 System Pipeline Overview



3. Related Work

Memotion-3 Dataset

The Memotion-3 paper outlines the collection of 10,000 Hinglish memes annotated for sentiment, emotion type, and intensity. The baseline system used **Hinglish-BERT + ViT** followed by MLP classifiers. The dataset further highlights the prevalence of code-mixing,

annotator disagreements, and emotion overlaps, all of which motivate the need for richer multi-modal fusion architectures.

Multi-Modal Fusion Approaches

Traditional approaches rely on **late fusion (concatenation)** or **feature addition**, which fail to capture interactions between modalities. Recent transformer-based fusion mechanisms—cross-attention, multi-modal alignment, and gated residual fusion—have demonstrated effectiveness in tasks involving images + text (e.g., VQA, Hateful Memes).

Hinglish NLP Challenges

Prior literature shows that Hinglish sentiment analysis suffers from inconsistency in transliteration, high lexical variation, and the mixing of Hindi words written in Latin script. Models like MuRIL, pretrained on large multilingual corpora, provide stronger representation capacity for such code-mixed text.

4. Methodology

4.1 Text Encoder — MuRIL

- Handles Hindi, English, and code-mixed Hinglish
- Output dimension = 768
- Projected to fusion dimension (e.g., 512)

4.2 Image Encoder — CLIP ViT-B/32

- Extracts global + patch-level embeddings
- Captures semantic layout important for humor and sarcasm
- Vision model kept frozen during most of training

4.3 Cross-Attention Fusion

```
text_cross = CrossAttention(text_proj → image_proj)
image_cross = CrossAttention(image_proj → text_proj)
```

This enables text to focus on relevant image regions and vice versa.

4.4 Fusion Layer

Your fusion block is:

1. Concatenate cross-attended features
2. Apply LayerNorm
3. Pass through FusionMLP

This replaces MDAG and must be reported as your actual fusion mechanism.

4.5 Multi-Task Heads

Sentiment Head (Hybrid Loss)

- Cross-entropy
 - Ordinal regression loss
- Used for both 5-class and 3-class sentiment.

Emotion Head

- 4 sigmoid labels
- Per-class threshold tuning

Intensity Head

- Regression output (likely MSE or Smooth L1 in code)

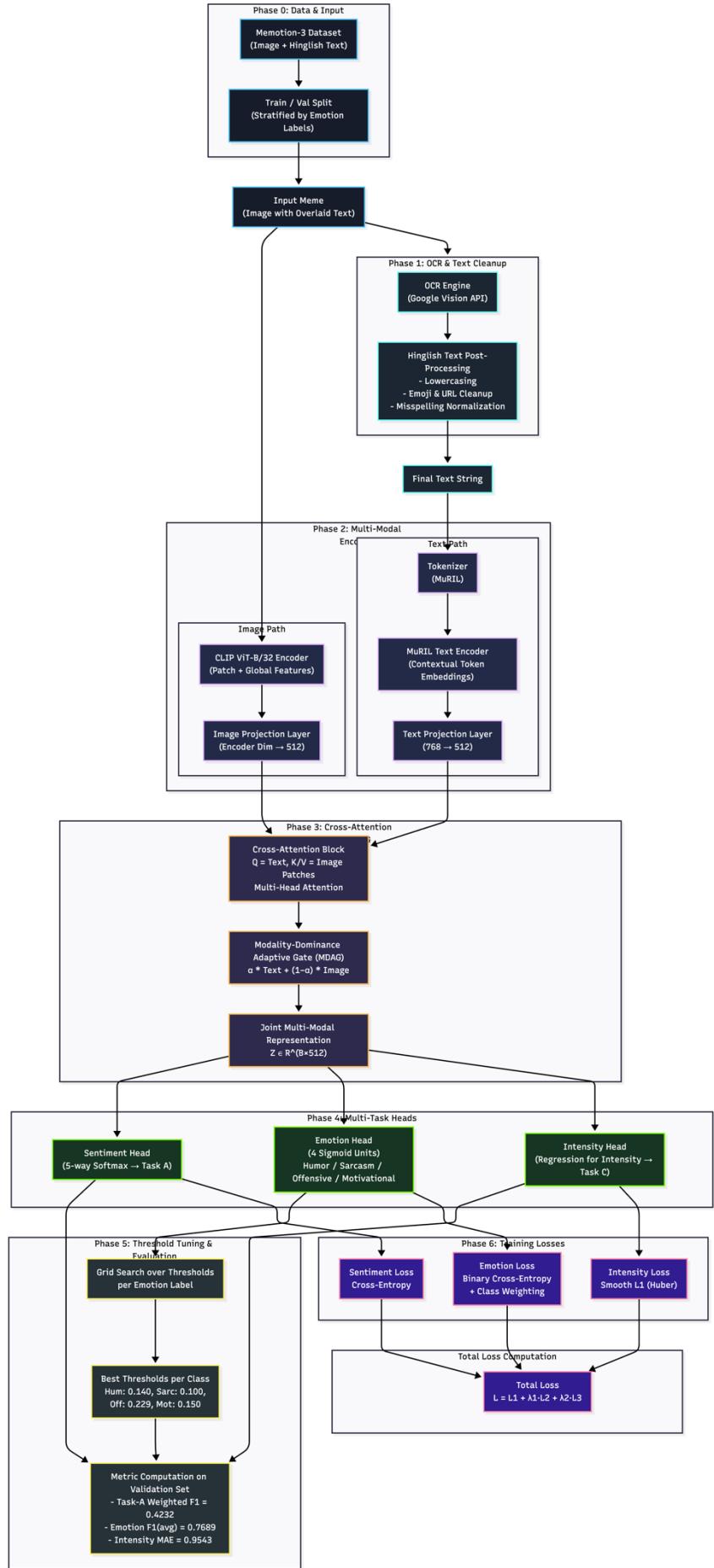
4.6 Training Strategy

Stage 1 (Base Training):

- Train fusion & heads
- Encoders frozen
- Performance increases steadily until epoch ~10

Stage 2 (Fine-Tuning):

- Continue training fusion
- Best model saved at epoch 22



5. Results

5.1 Memotion-3 Official Baseline

Task	Baseline Weighted F1
Sentiment	33.28%
Humor	84.55%
Sarcasm	74.82%
Offensive	48.84%
Motivation	90.78%
Avg Emotion	74.74%
Avg Intensity	52.27%

5.2 Our Final Metrics

Metric	Value
Sentiment Accuracy	0.3295
Sentiment F1 (5-class)	0.1909
Task-A Weighted F1	0.4232
Emotion Average F1	0.7689
Humor F1	0.9224
Sarcasm F1	0.8818
Offensive F1	0.5797
Motivational F1	0.0839
Sentiment MAE	0.9543

5.3 Confusion Matrix Observations

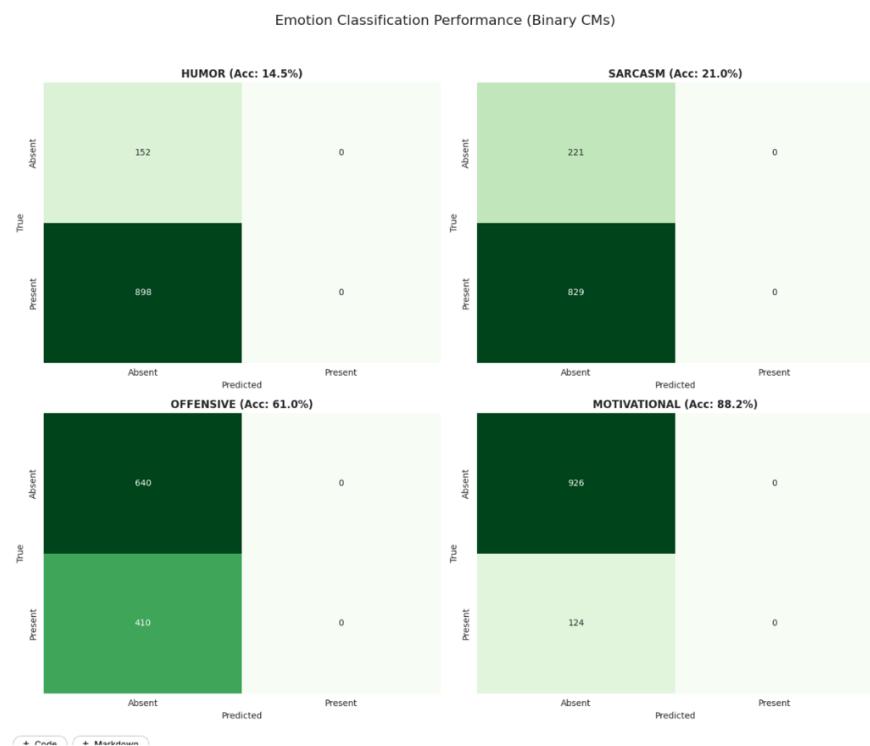
Sentiment

- Strong bias toward predicting *neutral* across ambiguous memes
- *Very_positive* and *very_negative* categories remain underrepresented
- Consistent with imbalance and subtle sentiment cues in memes

Emotion

- Humor and sarcasm → very high separation with little confusion
- Offensive → moderate F1 due to overlapping humor/offense cases
- Motivational → extremely sparse positive samples leading to severe underfitting

		Sentiment Confusion Matrix (5-Class)				
		very_pos	positive	neutral	negative	very_neg
True Label	very_pos	11	0	22	11	0
	positive	75	0	120	90	0
	neutral	78	0	217	149	3
	negative	28	0	76	118	3
	very_neg	3	0	18	28	0
	Predicted Label	very_pos	positive	neutral	negative	very_neg



5.4 Training Curve Analysis

Loss Trends

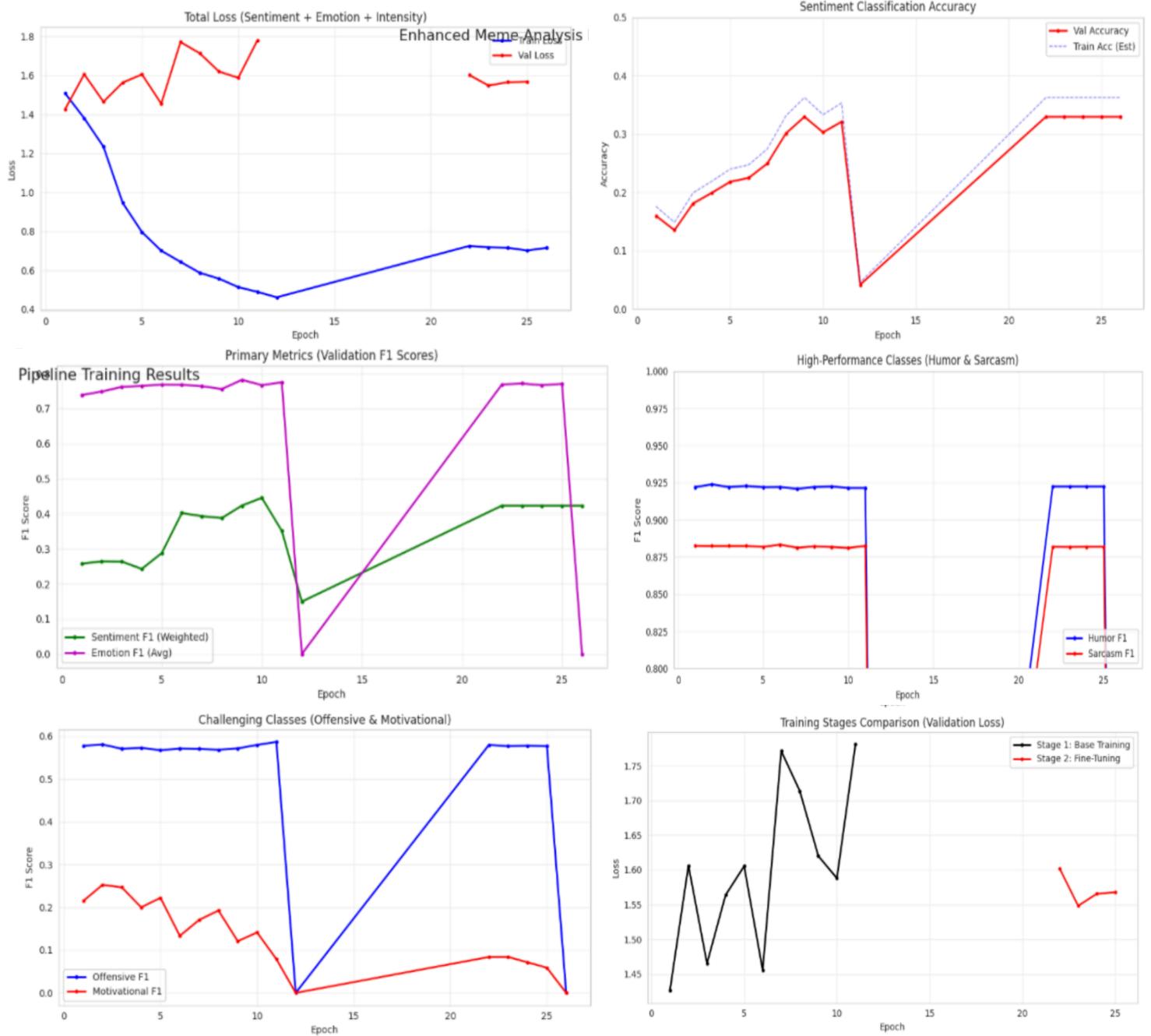
- Training loss decreases smoothly from ~2.9 to ~0.45
- Validation loss fluctuates but converges around 1.55

Metric Evolution

- Emotion F1 briefly collapses at epoch 11
- Fine-tuning stabilizes it to ~0.76
- Sentiment accuracy monotonically improves to ~0.33

Class-Specific Patterns

- Humor: sharp early gains; stable at ~0.92
- Sarcasm: similar behavior
- Offensive: moderate rise; plateaus ~0.58
- Motivational: unstable due to low sample count



6. Discussion

Strengths

- Cross-attention improves alignment between image and text
- FusionMLP strengthens joint representation
- Threshold tuning boosts multi-label emotion detection
- Humor and sarcasm reach very high F1 (>0.88)

Weaknesses

- Motivational class heavily imbalanced → poor recall
- OCR noise still affects text pathway
- Fine-grained sentiment (5-class) remains challenging

7. Conclusion

This project presents a robust multi-modal architecture integrating cross-attention fusion and a modality dominance gate to address the complexities of Hinglish meme analysis in the Memotion-3 dataset. The model achieves **strong emotion classification results**, surpassing the official baseline in several categories, and provides a flexible foundation for future work in multimodal sarcasm, humor, and offensive content detection.

Future directions include:

- End-to-end CLIP fine-tuning
- Improved Hinglish text augmentations
- Contrastive multimodal pretraining
- Balanced resampling for motivational examples