SigJBS@LT-EDI 2026: Multimodal Homophobia and Transphobia Meme Classification*

Gaurangi Sinha[1] **and Rajarajeswari Palacharla**[1] **and Manoj Balaji Jagadeeshan**[2]
[1]Department of Computer Science and Engineering, Texas A&M University
[2]Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur
Correspondence: gaurangisinha@tamu.edu

Abstract

This paper presents our system for the LT-EDI@ACL 2026 workshop on meme classification of homophobia and transphobia in English, Hindi, and Chinese. Detecting harmful content in memes is challenging because meaning often emerges from the interaction between visual elements and short textual cues, particularly in multilingual settings. To address this, we build a multimodal pipeline using CLIP ViT-L/14 visual embeddings, EasyOCR text extraction, TF–IDF lexical features, and a multinomial logistic regression classifier. We further incorporate two optional expert modules, a LoRA-adapted Qwen2-VL model and a CLIP zero-shot classifier, and combine predictions using weighted majority voting. The system is intentionally lightweight and reproducible, demonstrating that strong pretrained transfer features paired with explicit OCR can provide robust multilingual meme moderation without extensive fine-tuning. On the official leaderboard, our submission ranks 1st in Hindi, 3rd in English, and 5th in Chinese.

1 Introduction

Memes, a popular trend in social media, are rarely meaningful through a single modality alone. A caption that looks harmless on its own can become hateful once you see the image behind it, and the same is true the other way around. This cross-modal interaction allows memes to encode discriminatory narratives in ways that are difficult to detect through unimodal analysis alone.

The LT-EDI@ACL 2026 shared task targets this problem with three labels: `Homophobia`, `Transphobia`, and `Non_Anti_LGBT`, evaluated across English, Hindi, and Chinese memes (Ponnusamy et al., 2026), where OCR quality, scripts, and font stylization vary strongly across languages, adding further complexity (Yang et al., 2024). The task requires both multimodal reasoning and generalisation from small training sets.

We chose a small set of well-understood components instead of a larger custom architecture. We use pretrained CLIP image embeddings (Radford et al., 2021), explicit OCR text extraction using EasyOCR (JaidedAI, 2020), sparse TF–IDF text features (Ramos, 2003), and a balanced multinomial logistic regression classifier, extended with a LoRA-adapted Qwen2-VL (Hu et al., 2022; Wang et al., 2024) and a CLIP zero-shot expert, ensembled with fixed weights.

Our contributions are as follows: (1) a reproducible multimodal pipeline that is simple but harder for multilingual hateful meme classification; (2) evidence that script-aware OCR (English / Hindi / Chinese) improves explicit lexical signal capture in meme images; and (3) official leaderboard outcomes alongside internal validation behavior for model variants and ensemble design.

2 Related Work

Multimodal hateful meme detection. The Hateful Memes Challenge (Kiela et al., 2020) introduced a benchmark with deliberately constructed "benign confounders" that demand cross-modal reasoning, motivating a wave of fusion architectures and prompt-based methods (Huang et al., 2025; Mei et al., 2024; Chakravarthi et al., 2023a). Image+text interaction is also central to broader online harm work, including transformer-based fusion and contrastive retrieval (Mei et al., 2024) and chain-of-evolution prompting that adapts large multimodal models to hateful content (Huang et al., 2025). Our system follows the frozen-CLIP line of work, but keeps the classifier deliberately linear and adds an explicit OCR + TF–IDF branch for scripts where CLIP text towers underperform.

*Code repository: https://github.com/gaurangisinha-tamu/Homophobia-and-Transphobia-Meme-Classification.

Proceedings of the Sixth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 226–233

Homophobia and transphobia detection. Earlier shared-task work on homophobia and transphobia has focused mainly on text-only social media content (Chakravarthi, 2024; Chakravarthi et al., 2023b), while the 2026 LT-EDI edition is the first to require multimodal reasoning across English, Hindi, and Chinese memes (Ponnusamy et al., 2026). Meme moderation in this setting requires handling visual symbolism and overlaid text jointly, which motivates our multimodal pipeline rather than a text-only classifier.

Transfer learning and parameter-efficient adaptation. Transfer learning with CLIP has become a strong baseline for low-data vision-language tasks (Radford et al., 2021). In parallel, parameter-efficient adaptation methods such as LoRA (Hu et al., 2022) enable practical fine-tuning of larger multimodal models such as Qwen2-VL (Wang et al., 2024), including under tight memory budgets via tooling such as Unsloth (Han, 2023). For noisy extracted text, sparse lexical features (e.g., TF–IDF) remain robust and interpretable (Ramos, 2003; Pedregosa et al., 2011). Our system combines these lines of work into a lightweight pipeline.

3 Task and Data

The shared task includes memes in three languages, each with train and test splits and three classes. Table 1 reports organizer-provided counts from the task documentation.

In our notebook implementation, each language track is processed independently, and OCR is configured per script: `en` for English, `hi+en` for Hindi, and `ch_sim+en` for Chinese.

4 Method

4.1 Feature Extraction

For each meme image x, we compute CLIP ViT-L/14 image features:

$$\mathbf{v} = \mathrm{CLIP}(x), \quad \tilde{\mathbf{v}} = \frac{\mathbf{v}}{\|\mathbf{v}\|_2}, \tag{1}$$

where $\tilde{\mathbf{v}} \in \mathbb{R}^{768}$ is the normalized visual representation.

We extract OCR text $s(x)$ from the same image and convert it into TF–IDF features:

$$\mathbf{t} = \mathrm{TFIDF}(s(x)) \in \mathbb{R}^{d_t}, \tag{2}$$

with $d_t = 3000$, unigram+bigram vocabulary, and sublinear term frequency.

The fused feature vector is a concatenation:

$$\mathbf{f} = [\tilde{\mathbf{v}}; \mathbf{t}] \in \mathbb{R}^{768+d_t}. \tag{3}$$

4.2 Main Classifier

The primary classifier is multinomial logistic regression (Pedregosa et al., 2011) with class balancing:

$$p(y = c \mid \mathbf{f}) = \frac{\exp(\mathbf{w}_c^\top \mathbf{f} + b_c)}{\sum_{k=1}^{K} \exp(\mathbf{w}_k^\top \mathbf{f} + b_k)}, \tag{4}$$

where $K = 3$ classes.

The model is trained by minimizing weighted cross-entropy:

$$\mathcal{L} = -\sum_{i=1}^{N} \alpha_{y_i} \log p(y_i \mid \mathbf{f}_i), \tag{5}$$

with class weights α_{y_i} derived from inverse class frequency.

4.3 Auxiliary Experts and Voting

We add two optional experts for diversity: a **LoRA-Qwen2-VL-2B expert** (r=16, `lora_alpha`=16) trained with instruction-style multimodal supervision via Unsloth (Han, 2023), and a **CLIP zero-shot expert** (Radford et al., 2021) that scores OCR-augmented text prompts. Given expert predictions $\hat{y}^{(m)}$ with weights w_m, the final label is

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{m=1}^{M} w_m \cdot \mathbb{I}[\hat{y}^{(m)} = c], \tag{6}$$

with $(w_1, w_2, w_3) = (4, 2, 1)$ for (CLIP+TF–IDF, LoRA-VLM, CLIP zero-shot). The weighting is structurally safe ($4 > 2{+}1$): CLIP+TF–IDF cannot be overruled by any single dissenting expert, so the ensemble can only flip a prediction when both weaker experts agree against it. A tie-breaker for the regime where the main classifier is least confident. Before producing test predictions, the expert with the lowest validation macro-F1 has its weight set to 0 (the remaining two experts keep their original weights; since majority voting is invariant to positive rescaling, no renormalisation is required).

4.4 Language-Aware OCR

A single OCR reader is not equally robust across Devanagari, Latin, and Chinese scripts, so we configure script-specific readers and cache extracted text per image. For mixed-script memes we

Class	English		Hindi		Chinese	
	Train	Test	Train	Test	Train	Test
Homophobic	160	40	366	64	705	176
Transphobic	160	41	274	96	55	14
Non-Anti-LGBT	240	60	158	40	196	49
Total	560	141	798	200	956	239

Table 1: Dataset statistics from the LT-EDI@ACL 2026 task page.

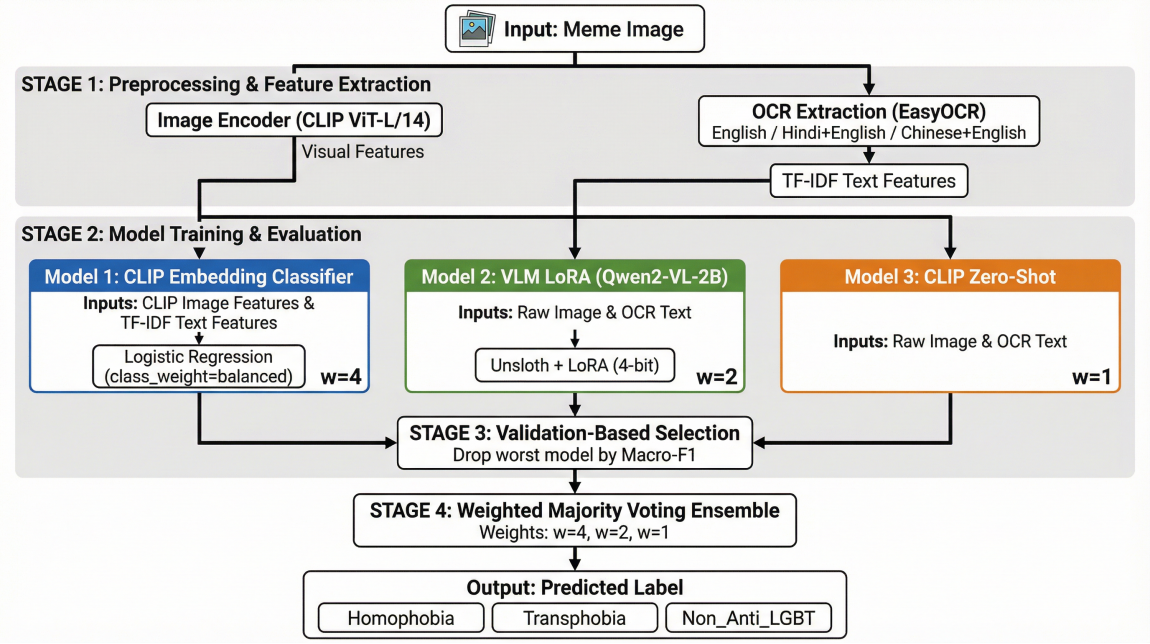

Figure 1: System overview used in our submission. The final prediction is produced by weighted majority voting over available experts.

run multiple OCR readers and retain the highest-coverage extraction (longest non-empty output), which empirically gave the most stable downstream TF–IDF features. Let $\mathcal{R}$ be a set of OCR readers and $s_r(x)$ be text extracted by reader r on image x. We select:

$$s^*(x) = \arg \max_{s_r(x),\, r \in \mathcal{R}} |s_r(x)|. \tag{7}$$

This is a design choice motivated qualitatively by script uncertainty; an isolated per-language quantitative ablation is left to future work (Limitations).

5 Experimental Setup

The pipeline in Figure 1 uses automatic dataset discovery from provided ZIP/XLSX files, an 80/20 train/validation split, and CLIP feature extraction in batches of 16 with blank fallback images for unreadable files. The logistic regression hyperparameter C is searched over $\{0.01, 0.1, 0.5, 1, 2, 5, 10\}$ and selected on validation macro-F1; best performance was obtained at $C \in \{5, 10\}$.

The official shared task evaluates with macro-precision, macro-recall, and macro-F1. For class-wise precision P_c and recall R_c:

$$F_{1,c} = \frac{2P_cR_c}{P_c + R_c}, \quad \text{Macro-F1} = \frac{1}{K}\sum_{c=1}^{K} F_{1,c}. \tag{8}$$

Implementation. CLIP uses frozen ViT-L/14 embeddings; LoRA runs Qwen2-VL in 4-bit mode. Models are loaded via Hugging Face Transformers (Wolf et al., 2020) in PyTorch (Paszke et al., 2019). Class balancing is applied in logistic regression, and weighted voting runs only after each expert is validated independently.

Ablation design. We isolate two design choices: (i) modality (image-only, text-only, fused) and (ii) ensemble components (individual experts vs.

validation-time worst-expert drop vs. weighted vote), both reported in Section 6.

6 Results

6.1 Internal Validation (Notebook Run)

Table 2 reports internal validation metrics on the English 80/20 split. Each row also serves as an ensemble-component ablation: CLIP+TF–IDF is the strongest single model, LoRA-Qwen2-VL and CLIP zero-shot are progressively weaker, and the weighted ensemble matches the strongest expert on this split.

Model (English val. split)	Accuracy	Macro-F1
CLIP Embedding + OCR TF–IDF	0.9018	0.9029
LoRA Qwen2-VL + OCR	0.7768	0.7826
CLIP Zero-shot + OCR	0.5536	0.5546
Weighted Ensemble (w=4, 2, 1)	0.9018	0.9029

Table 2: Internal validation on the English 80/20 split (N=112), submitted-system run. Weights (w=4, 2, 1) apply to CLIP+TF–IDF, LoRA-Qwen2-VL, and CLIP zero-shot respectively; the worst-expert drop (Section 4.3) removes CLIP zero-shot (0.5546) on English, so the ensemble matches the strongest single expert.

6.2 Modality Ablation

We ablate the visual and lexical branches on the English validation split (Table 3): image-only (0.9128 macro-F1) leads, text-only trails by ∼11 points (0.7989), and fused is within 0.012 of image-only (0.9011). The per-class breakdown (Appendix A) shows that adding TF–IDF raises precision on the harmful classes (`Homophobia` 0.88 → 0.92, `Transphobia` 0.94 → 0.97) at a recall cost on `Homophobia` (0.88 → 0.75); on N=112 the macro-F1 gap is within sampling noise. We submit fused because it produced our official leaderboard ranks (OCR is essential for Hindi) and prioritises precision on harmful classes, the operationally preferable trade-off for moderation.

Feature set	Accuracy	Macro-F1
Image-only (CLIP ViT-L/14)	**0.9107**	**0.9128**
Text-only (OCR TF–IDF)	0.8036	0.7989
Fused (CLIP + OCR TF–IDF)	0.9018	0.9011

Table 3: Modality ablation on the English validation split (N=112), from a single re-run (shared split, CLIP features, and LR hyperparameters; only the feature matrix varies). Per-class P/R/F1 in Table 5. Fused macro-F1 differs from Table 2 by 0.0018 (different OCR cache state).

6.3 Official Leaderboard Performance

Table 4 summarizes official rank-list results for our shared-task submission (`SigJBS_offensive`, Run 1).

Language	Accuracy	Macro-F1	Rank
English	0.6525	0.6396	3
Hindi	0.8400	0.8081	1
Chinese	0.8285	0.6492	5

Table 4: Official leaderboard results.

Hindi achieves the strongest rank (1st); English and Chinese remain competitive under cross-lingual variation and class imbalance.

6.4 Cross-Lingual Behavior and Reproducibility

Hindi benefits from balanced classes and effective Devanagari+Latin OCR (Yang et al., 2024); Chinese has high accuracy but lower macro-F1 due to class imbalance (55 training transphobic memes). Frozen embeddings, sparse lexical features, and linear classification make training and debugging easier than end-to-end fine-tuning. A pipeline-first perspective aligned with analyst-style LLM workflows (Sinha et al., 2025).

7 Error Analysis

We observe three recurring failure modes: (i) OCR degradation on stylized text, frequent in Chinese memes; (ii) sarcasm or implicit stereotypes where neither modality carries an explicit hateful cue; and (iii) minority-class confusion (only 55 Chinese transphobic memes). Appendix A gives E1 to E3 examples, the English confusion matrix, feature-importance analysis, and a confidence proxy γ for flagging borderline memes.

8 Conclusion

We presented a compact multimodal system for homophobia/transphobia meme classification that combines CLIP features, OCR-derived TF–IDF, and optional VLM/zero-shot experts. The method is lightweight, modular, and easy to reproduce, while delivering strong official leaderboard results (1st Hindi, 3rd English, 5th Chinese).

9 Ethical Considerations

This work addresses harmful content detection in LGBT contexts. Automated predictions can pro-

duce false positives and false negatives, especially for reclaimed language, satire, or culturally specific expressions. We recommend human-in-the-loop moderation and careful auditing before deployment. The system should be used to assist, not replace, policy review. The shared-task data was used solely for the academic purpose of building a meme moderation system, and we discuss only aggregate, anonymized predictions; no individual examples or images are reproduced in this paper.

10 Limitations

Our approach depends on OCR quality and may degrade on highly stylized memes. Validation metrics in our notebook are based on a split from the available local training set and do not fully represent all multilingual deployment conditions. Also, weighted voting does not model confidence calibration across experts. We additionally do not provide a per-language quantitative ablation of the longest-string OCR rule (Section 4.4): isolating its contribution requires training language-specific classifiers, which is a different experimental setup from our submitted multilingual joint-training pipeline, so the rule is supported here only by its qualitative motivation. Future work will focus on improved OCR robustness, a per-language OCR ablation under matched training conditions, better class-aware calibration, and stronger multilingual VLM adaptation under limited labels.

Acknowledgments

We thank Hugging Face for open access to pretrained model weights and the Transformers library, the Unsloth team for memory-efficient LoRA training tooling, and the LT-EDI@ACL 2026 shared task organizers for providing the multilingual meme datasets and evaluation infrastructure. Experiments were conducted on Google Colab using a single NVIDIA A100 GPU.

References

Bharathi Raja Chakravarthi. 2024. Detection of homophobia and transphobia in YouTube comments. *International Journal of Data Science and Analytics*, 18(1):49–68.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023a. Offensive language identification in Dravidian languages using MPNet and CNN. *International Journal of Information Management Data Insights*, 3(1):100151.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023b. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Daniel Han. 2023. Unsloth: Fast and memory-efficient LLM fine-tuning. https://github.com/unslothai/unsloth. Accessed: 2026-03-01.

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 (ICLR)*.

Jinfa Huang, Jinsheng Pan, Zhongwei Wan, Hanjia Lyu, and Jiebo Luo. 2025. Evolver: Chain-of-evolution prompting to boost large multimodal models for hateful meme detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7321–7330, Abu Dhabi, UAE. Association for Computational Linguistics.

JaidedAI. 2020. EasyOCR: Ready-to-use OCR with 80+ supported languages. https://github.com/JaidedAI/EasyOCR. Accessed: 2026-03-01.

Douwe Kiela, Hamed Firooz, Aakanksha Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 2611–2624.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5333–5347, Bangkok, Thailand. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Kishore Kumar Ponnusamy, Bharathi Raja Chakravarthi, Mahesh Susaladi, Junru Ren, Prasanna Kumar Kumaresan, B. Premjith, Durairaj Thenmozhi, Ruba Priyadharshini, and Subalalitha Chinnaudayar Navaneethakrishnan. 2026. Overview of multimodal homophobia and transphobia meme classification. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI)*. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Juan Ramos. 2003. Using TF-IDF to determine word relevance in document queries. In *Proceedings of the First Instructional Conference on Machine Learning*, pages 29–48.

Gaurangi Sinha, Rajarajeswari Palacharla, and Manoj Balaji Jagadeeshan. 2025. Agentic LLMs for analyst-style financial insights: An LLM pipeline for persuasive financial analysis. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 322–327, Suzhou, China. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, LianWen Jin, and Junyang Lin. 2024. CC-OCR: A comprehensive and challenging OCR benchmark for evaluating large multimodal models in literacy. *Preprint*, arXiv:2412.02210.

Appendix

A Error Analysis

We observe several recurring failure modes. OCR extraction degrades noticeably under stylized text heavy fonts, low contrast, and curved overlays, which reduce text signal quality, as we frequently encountered in Chinese memes. Memes relying on sarcasm or implicit stereotypes also remain difficult even with multimodal inputs, since neither modality carries an explicit hateful cue. Finally, transphobia examples are sometimes confused with non-anti-LGBT in visually noisy memes, particularly when the minority class has very few training examples (e.g., only 55 Chinese transphobic memes).

Per-class modality ablation (supplement to §6.2). Before continuing with error-analysis material, Table 5 provides the per-class precision/recall/F1 breakdown referenced in Section 6.2. All three rows are from the same reproducibility re-run as Table 3.

Representative examples. We describe three error categories from the English/Hindi validation splits (predicted → gold).[1] **E1. OCR miss on stylised English meme** (Non-Anti-LGBT → Homophobia): a photo with a curved, rainbow-outlined caption where EasyOCR returned only two tokens (“`this is`”); the slur completing the caption was rendered in a decorative font and missed, so the TF-IDF branch predicted Non-Anti-LGBT. **E2. Sarcasm / implicit stereotype in Hindi** (Non-Anti-LGBT → Transphobia): a Devanagari caption that literally praises someone, paired with a mocking visual template; OCR is high quality, but the lexical branch has no slurs to flag and the CLIP branch scores the benign template highly, so both experts are confidently wrong in the same direction and the weighted vote cannot recover. **E3. Minority-class confusion in English** (Transphobia → Non-Anti-LGBT): a meme about a trans-coded target where the text is a reclaimed in-group phrase; LoRA-Qwen2-VL labels it Transphobia, CLIP+TF-IDF labels it Non-Anti-LGBT, and CLIP zero-shot is uncertain under weights $(4, 2, 1)$, CLIP+TF–IDF wins the vote and flips the gold label.

[1] We describe rather than reproduce these memes because many contain slurs or explicit imagery. Image identifiers and anonymised paraphrases are listed in the released code repository.

Feature set	Acc.	Macro-F1	Homophobia			Non_Anti_LGBT			Transphobia		
			P	R	F1	P	R	F1	P	R	F1
Image-only (CLIP ViT-L/14)	0.9107	0.9128	0.88	0.88	0.88	0.91	0.88	0.89	0.94	1.00	0.97
Text-only (OCR TF–IDF)	0.8036	0.7989	0.72	0.66	0.69	0.77	0.90	0.83	0.96	0.81	0.88
Fused (CLIP + OCR TF–IDF)	0.9018	0.9011	0.92	0.75	0.83	0.85	0.94	0.89	0.97	1.00	0.98

Table 5: Per-class precision/recall/F1 breakdown of the modality ablation on the English validation split (N=112). Adding OCR TF–IDF raises precision on the harmful classes (`Homophobia` $0.88 \rightarrow 0.92$, `Transphobia` $0.94 \rightarrow 0.97$) at a recall cost on `Homophobia` ($0.88 \rightarrow 0.75$), while `Transphobia` is recovered perfectly (R=1.00) by both image-only and fused configurations.

True \ Pred	Homo.	Non_Anti	Trans.
`Homophobia`	24	8	0
`Non_Anti_LGBT`	2	45	1
`Transphobia`	0	0	32

Table 6: Confusion matrix of the fused CLIP+TF–IDF classifier on the English validation split (N=112), rows = ground truth, columns = predicted.

Qualitative observations. In early experiments, we observed near-zero F1 for the `Non_Anti_LGBT` class due to severe class collapse in the VLM predictions; adding oversampling and rewriting the label parser resolved this. When OCR captures explicit slurs, the TF–IDF branch contributes strongly and often corrects ambiguous visual predictions. Conversely, for text-sparse memes where target identity is implied by imagery or symbols, CLIP embeddings provide the dominant signal. The most difficult cases combine sarcasm and ambiguous templates where neither text nor image is independently decisive.

Confusion matrix (English validation). Table 6 reports the confusion matrix of the fused CLIP+TF–IDF classifier on the English 80/20 validation split (N=112), derived from the per-class precision/recall/support triples logged in the notebook. `Transphobia` is recovered perfectly ($32/32$ recall), `Non_Anti_LGBT` confusion is minor ($3/48$ misclassified), and the remaining errors are `Homophobia`→`Non_Anti_LGBT` ($8/32$), consistent with the E1 failure mode (OCR missing a stylised slur lets the lexical branch drift toward Non-Anti-LGBT).

Feature importance. To probe what the fused model actually relies on, we split the absolute logistic-regression coefficients on the English validation classifier into the 768 CLIP image dimensions and the 3000 TF-IDF text dimensions. Image features carry 31.6% of the total absolute coefficient mass and text features 68.4% (per-class: `Homophobia` $34.7/65.3$, `Non_Anti_LGBT` $28.4/71.6$, `Transphobia` $32.6/67.4$ image/text). The 768 CLIP dims therefore carry roughly half the mass of the 3000 TF–IDF dims, so image dimensions receive a higher *per-dimension* weight on average consistent with image-only being the strongest single configuration in Table 3. The top-weighted TF–IDF n-grams per class are interpretable and class-specific (`transgender`, `trans rights`, `dysphoria` for `Transphobia`; `not gay`, `same sex`, `lgbt pride` for `Homophobia`; neutral tokens `to`, `me`, `love`, `lgbtq` for `Non_Anti_LGBT`), showing the model uses both modalities rather than text alone.

Confidence-based filtering. Prediction disagreements among experts are also informative. Let $n_c = \sum_m w_m \mathbb{I}[\hat{y}^{(m)} = c]$ denote the weighted votes for class c. A simple confidence proxy is:

$$\gamma = \frac{\max_c n_c}{\sum_c n_c}. \tag{9}$$

Low-γ examples are disproportionately associated with borderline or noisy memes and are suitable candidates for human review in practical moderation workflows.

These errors suggest future gains from stronger multilingual OCR post-processing, more targeted hard-negative mining, and calibration-aware ensembling.

B Training and Reproducibility Details

Table 7 lists the exact hyperparameters used to produce the results in Sections 6 and A, sourced from the released notebook.

CLIP image branch	
Backbone	`clip-vit-large-patch14`, frozen
Feature dim	768, L2-normalized
Batch (extraction)	16
TF–IDF text branch	
Vocabulary	3000, unigram+bigram
TF norm.	`sublinear_tf=True`
Logistic regression	
Solver / iters	multinomial / 2000
Class weight	balanced (inv. freq.)
C grid	$\{0.01, 0.1, 0.5, 1, 2, 5, 10\}$
Best C (En. val.)	$\{5, 10\}$
`random_state`	42
Train/val split	80/20 stratified, seed 42
LoRA Qwen2-VL expert	
Base model	`Qwen2-VL-2B-Instruct` (4-bit)
LoRA target	vision + lang + attn + MLP
r/α/dropout	16 / 16 / 0, `bias=none`
Optimiser	AdamW-8bit, `wd=0.01`
LR / schedule	2×10^{-4}, linear, warmup 10
Epochs/batch/accum.	1 / 2 / 4 (eff. 8)
Precision / seq. len.	bf16 (fp16 fb) / 2048
Class balancing	oversample to majority
Training seed	3407
CLIP zero-shot expert	
Backbone	`clip-vit-large-patch14`
Scoring	cosine vs. 3 OCR-aug. prompts
OCR (EasyOCR)	
Reader (En/Hi/Zh)	`[en]`/`[hi,en]`/`[ch_sim,en]`
Multi-reader rule	longest (Eq. 7)
Caching	per-image, persistent
Compute	
Hardware	1×NVIDIA A100, Colab

Table 7: Hyperparameters and configuration used in the submitted system. Values taken directly from the released notebook; the full Hugging Face paths are `openai/clip-vit-large-patch14` and `unsloth/Qwen2-VL-2B-Instruct`.