IndicHM: A Multilingual Benchmark and Disagreement-Centric Framework for Hateful Meme Detection

Shahid Shafi Dar[1], **Mohd Aamir**[1], **Arindol Sarkar**[2], **Hrishiraj Chowdhury**[2], **Pranjal Pandey**[3], **Nagendra Kumar**[1,*]
[1]Indian Institute of Technology Indore, Indore, India
[2]Rajiv Gandhi Institute of Petroleum Technology, Uttar Pradesh, India
[3]Indian Institute of Information Technology Bhagalpur, Bihar, India
{phd2201201004, mt2402101005, nagendra}@iiti.ac.in,
{24cd3007, 24cs2018}@rgipt.ac.in, pranjal.230103027@iiitbh.ac.in
*Corresponding author

Abstract

Hateful memes are difficult to detect because their harmful meaning often emerges from the interaction between image and text rather than from either modality alone. This challenge becomes more severe in multilingual settings, where memes frequently contain code-mixed language and culturally grounded references. Existing Indic hateful meme datasets lack scale, language coverage, or task diversity, and current multimodal methods often overlook cross-modal semantic divergence. We introduce **IndicHM**, a trilingual hateful meme benchmark containing over 15,004 manually annotated image–text pairs in English, Hindi, and Bengali, including code-mixed content, supporting both hate detection and target classification. We also propose **IndicHMNet**, a disagreement-centric framework that decomposes representations into shared agreement and modality-specific residual subspaces and propagates the resulting disagreement-aware features through multimodal fusion and contrastive representation learning. The proposed framework adaptively emphasizes modality-specific semantic divergence during fusion and optimization, enabling improved modeling of complex hateful meme semantics. Experiments across all languages and tasks demonstrate consistent gains over strong multimodal and vision-language baselines. The code and dataset are available on https://github.com/Shahid-135/Indic-Hateful-Meme-Dataset.

Disclaimer. This paper contains discriminatory and offensive content due to the nature of the task and dataset. Reader discretion is advised.

1 Introduction

Hateful memes are challenging to moderate because their harmful intent often emerges from interactions between image and text rather than from either modality alone (Kiela et al., 2020; Hee et al., 2023). A seemingly benign image paired with a hateful caption may appear harmless independently, yet becomes offensive when interpreted jointly. This dependence on cross-modal reasoning makes hateful meme detection fundamentally different from conventional text-only hate speech analysis. The problem becomes even more challenging in multilingual settings, where online discourse frequently involves code-mixed and low-resource languages (Singh et al., 2024; Hossain et al., 2022, 2024). Users commonly blend English with Hindi or Bengali, switch scripts, and employ culturally grounded references, sarcasm, and visual metaphors that are difficult to interpret using lexical features alone. Recent studies further show that multilingual hateful meme understanding requires joint reasoning across linguistic, cultural, and visual contexts (Bui et al., 2025; Kumari et al., 2024). Figure 1 shows some hateful meme examples.



(a) Gender Hate (b) Political Hate

(c) Religious Hate (d) Racism Hate

Figure 1: Representative hateful memes from the proposed dataset across four categories (Source: Reddit).

Existing hateful meme benchmarks have progressively expanded multimodal hate understanding across diverse linguistic settings. FHM (Kiela et al., 2020) introduced multimodal hate detection through benign confounders requiring joint image–

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 39910–39931
October 24-29, 2026 ©2026 Association for Computational Linguistics

text reasoning. MUTE (Hossain et al., 2022) and MIMIC (Singh et al., 2024) later extended the task to Bengali and Hinglish code-mixed memes, highlighting the challenges of multilingual and low-resource hate expression. CM-Off-Meme (Kumari et al., 2024), BHM (Hossain et al., 2024), and Multi^{3}Hate (Bui et al., 2025) further explored multilingual, code-mixed, and target-aware hate understanding. More recently, PrideMM (Shah et al., 2024) introduced fine-grained annotations for hate, stance, humor, and target categories. Despite these advances, existing benchmarks remain limited in multilingual scale, code-mixed coverage, or dual-task supervision. In contrast, IndicHM provides large-scale multilingual hateful meme annotations across English, Hindi, and Bengali with supervision for both hateful intent detection (Hate vs Non-Hate) and target classification. As summarized in Table 1, no existing benchmark simultaneously provides large-scale multilingual coverage, code-mixed data, and dual-task supervision across these Indic languages.

Beyond dataset limitations, existing multimodal methods are constrained by how they model modality interaction. Most approaches rely on concatenation or symmetric fusion (Li et al., 2019; Radford et al., 2021), implicitly assuming semantic agreement between image and text. However, hateful memes are often characterized by modality-specific semantic divergence, where image and text individually appear benign but jointly convey harmful intent. Moreover, vision-language embedding spaces exhibit modality gaps and anisotropic geometry (Liang et al., 2022), making direct feature comparison unreliable. Recent studies further show that disagreement and incongruity signals are informative for sarcasm and offensive meme understanding. A principled framework should therefore explicitly model modality-specific semantic divergence and propagate the resulting signal through multimodal fusion and representation learning.

In this work, we define cross-modal disagreement as modality-specific residual semantic information that cannot be explained within a shared image-text agreement subspace. Intuitively, hateful memes often contain visual or textual features whose semantics remain weakly aligned across modalities, even when each modality appears independently benign. Modeling these residual divergences enables the framework to preserve informative multimodal inconsistencies that are typically suppressed by agreement-focused fusion strategies.

Dataset	Languages	Code-Mixed	Size
FHM	EN	✗	10K
MUTE	BN	✓	4.1K
MIMIC	HI	✓	5.0K
Da-Meme	TA, ML	✗	1.7K
HatReD	EN	✗	3.2K
PrideMM	EN	✗	5.1K
BHM	BN	✓	7.5K
Multi^{3}Hate	EN, ES, DE, ZH, HI	✗	1.5K
IndicHM	**EN, HI, BN**	✓	**15K**

Table 1: Comparison of **IndicHM** with representative hateful meme datasets.

To address these challenges, we introduce **IndicHM**, a large-scale trilingual hateful meme benchmark containing over 15K manually annotated image–text pairs across English, Hindi, and Bengali, including code-mixed variants. The dataset supports both hateful intent detection and target classification across religion, politics, gender, and race. We further propose **IndicHMNet**, a disagreement-centric multimodal framework that models hateful memes through shared agreement and modality-specific residual semantics. IndicHMNet decomposes representations into shared agreement and residual subspaces, and propagates the resulting disagreement-aware signal through multimodal fusion, topology-aware contrastive learning, and prototype alignment. This design enables semantically divergent memes to exert stronger influence during optimization while preserving modality-specific multimodal cues that are typically suppressed by conventional fusion strategies. The main contributions are as follows:

- We introduce **IndicHM**, a large-scale trilingual hateful meme benchmark containing over 15K manually annotated image–text pairs across English, Hindi, and Bengali, including code-mixed content with supervision for hateful intent and hate target classification.
- We propose **IndicHMNet**, a disagreement-centric framework that decomposes image and text representations into shared agreement and modality-specific residual subspaces, and propagates the resulting disagreement-aware signal across multimodal fusion, contrastive learning, and prototype alignment.
- We benchmark IndicHMNet against unimodal, multimodal, and recent vision-language baselines across all languages and tasks and provide ablation studies demonstrating the effectiveness of disagreement-centric learning.

2 Related Work

Existing work can broadly be grouped into two directions: multimodal hateful meme benchmarks and multilingual code-mixed moderation.

Multimodal Hateful Meme Benchmarks: The field gained significant attention with the Facebook Hateful Memes (FHM) benchmark (Kiela et al., 2020), which introduced benign confounders to encourage multimodal reasoning beyond unimodal learning. Subsequent datasets expanded this direction toward multilinguality, explainability, and fine-grained hate analysis, including HatReD (Hee et al., 2023), Multi^{3}Hate (Bui et al., 2025), MUTE (Hossain et al., 2022), BHM (Hossain et al., 2024), and MIMIC (Singh et al., 2024). Recent benchmarks such as CM-Off-Meme (Kumari et al., 2024) and Memotion 3 (Mishra et al., 2023) further explored offensiveness, implicit intent, sentiment, and emotion in code-mixed meme settings. More recently, PrideMM (Shah et al., 2024) extended meme understanding beyond binary hate classification through multi-attribute annotations for hate, target, stance, and humor in LGBTQ+ Pride-related memes. However, existing multilingual hateful meme datasets remain limited in scale, language diversity, or task coverage. IndicHM addresses these limitations through a large-scale trilingual benchmark supporting both hateful meme detection and hate target classification.

Multilingual and Code-Mixed Moderation: Multilingual and code-mixed hate detection remains challenging due to cross-lingual transfer gaps, implicit semantics, and culturally grounded expressions (Bui et al., 2025; Mathew et al., 2021; Pavlopoulos et al., 2021). In South Asian social media, code-mixing between English and Indic languages is often accompanied by transliteration, script switching, and contextual references. Prior studies also show that multilingual moderation systems struggle to generalize across culturally diverse and low-resource settings, particularly for implicit and multimodal hate. Benchmarks such as CM-Off-Meme (Kumari et al., 2024), Memotion 3 (Mishra et al., 2023), MUTE (Hossain et al., 2022), and MIMIC (Singh et al., 2024) highlight these challenges in multilingual meme settings. These findings motivate multilingual multimodal datasets and architectures capable of modeling culturally grounded meme discourse. Additional related works are discussed in Appendix A.

3 Indic Hate Meme Dataset

We introduce **IndicHM**, a large-scale multilingual benchmark for hateful meme understanding in the Indian social media ecosystem. IndicHM spans three major languages, such as English, Hindi, and Bengali, and contains both monolingual and code-mixed memes, reflecting realistic communication patterns commonly observed in Indic online discourse. Unlike prior hateful meme datasets that primarily focus on a single language or limited linguistic diversity, IndicHM provides a unified benchmark supporting multilingual evaluation under consistent task definitions. The dataset supports both hateful meme detection and fine-grained target classification across diverse sociocultural contexts.

3.1 Dataset Statistics

IndicHM contains 15,004 memes evenly distributed across English, Hindi, and Bengali. Table 2 summarizes the label distributions for both tasks. For Task 1, the dataset maintains a near-balanced distribution with 7,504 hateful and 7,500 non-hateful memes. For Task 2, hateful memes are categorized into five target classes: Religion, Gender, Politics, Race or Ethnicity, and Other. The dataset also contains substantial code-mixed and transliterated content, particularly in Hindi and Bengali subsets, reflecting real-world multilingual social media discourse. Extended dataset statistics are provided in Appendix B.

3.2 Dataset Collection and Filtering

IndicHM was constructed through a multi-stage collection, filtering, and annotation pipeline designed to obtain diverse and representative memes while reducing collection and annotation biases. We initially collected approximately 45,000 publicly available memes from Reddit and Pinterest using neutral and broadly defined search keywords, including "Bengali Troll Memes," "English Troll Memes," "Funny Memes," and "Racist Memes." Importantly, the collection process did not target specific hate categories or individual communities, thereby reducing the risk of introducing category-specific sampling bias.

The collected memes were subsequently processed through several filtering stages. First, low-quality, corrupted, irrelevant, and visually unusable samples were removed. Second, the language and script of the textual content were verified to ensure that the samples belonged to the target language

Language	Task–1: Binary		Task–2: Multiclass (Hate)					Total
	Hate	Non-Hate	Gender	Religion	Political	Racism	Others	
Hindi	2,509	2,500	680	608	560	223	438	5,009
English	2,495	2,500	810	695	409	332	249	4,995
Bengali	2,500	2,500	500	500	500	500	500	5,000
Total	7,504	7,500	1,990	1,803	1,469	1,055	1,187	15,004

Table 2: Dataset statistics for binary hate detection (Task 1) and hate target classification (Task 2).

Category	English	Hindi	Bengali
Hateful Intent	0.82	0.78	0.76
Religion	0.85	0.81	0.79
Gender	0.79	0.75	0.74
Race or Ethnicity	0.81	0.77	0.76
Political	0.78	0.79	0.77
Average	**0.81**	**0.78**	**0.76**

Table 3: Inter-annotator agreement (κ).

groups. Near-duplicate and duplicate memes were then identified and removed to prevent redundant samples and potential information leakage. Since conventional OCR systems often perform poorly on Indic scripts, we further employed human-in-the-loop transcription to accurately recover meme text, including transliterated and code-mixed content. The resulting samples were manually annotated by native speakers following a calibrated annotation protocol for reliable hateful/non-hateful classification.

Following annotation and quality verification, we curated the final IndicHM benchmark containing 15,004 memes. The final dataset was constructed to maintain a near-balanced distribution between hateful and non-hateful samples while preserving proportional representation across the included languages and data sources. The overall dataset construction process can be summarized as:

Raw Collection ($\sim$45K) $\rightarrow$ Quality Filtering $\rightarrow$ Language Verification $\rightarrow$ Near-Duplicate Removal $\rightarrow$ Human Transcription $\rightarrow$ Manual Annotation $\rightarrow$ Balanced Sampling $\rightarrow$ IndicHM (15,004)

This multi-stage procedure ensures that the final benchmark contains linguistically diverse, non-redundant, and manually verified multimodal samples suitable for evaluating hateful meme detection in Indic languages. We have also extended the dataset collection in the Appendix C.

3.3 Annotation Pipeline

The annotation process involved six native speakers, with two annotators assigned per language. All annotators were additionally fluent in English to ensure consistent interpretation of code-mixed content. Before full-scale annotation, annotators completed a calibration phase using a curated seed set of 200 memes containing challenging examples such as implicit hate, sarcasm, and culturally contextual references. Disagreements were periodically reviewed with two senior researchers, and annotation guidelines were refined iteratively throughout the process. Each meme was annotated for two tasks: (i) hateful intent detection and (ii) target classification. Task 1 consists of binary classification into hateful and non-hateful content. For memes identified as hateful, Task 2 assigns the primary hate target category from Religion, Gender, Race, Politics, and Other. Detailed annotation guidelines are provided in the Appendix D

3.4 Inter-Annotator Agreement

We evaluate annotation reliability using Cohen's κ, reported in Table 3. The average agreement across all languages is $\kappa \approx 0.78$, indicating substantial annotation consistency. English achieves the highest agreement (0.81), followed by Hindi (0.78) and Bengali (0.76). Agreement is strongest for Religion-focused hate, where harmful intent is often expressed explicitly through recognizable symbols or terminology. In contrast, Gender-based hate exhibits comparatively lower agreement, particularly in Bengali, due to its comparatively implicit and context-dependent nature.

4 Methodology

Hateful memes often express harmful intent through semantic inconsistency between image and text rather than semantic agreement alone. To model this behavior, we propose **IndicHM-Net**, a disagreement-centric framework consisting of: (i) Cross-Modal Disagreement Projection (CMDP), (ii) Asymmetric Disagreement-Aware Fusion (ADAF), (iii) Disagreement-Weighted Adaptive Topology-aware Contrastive Learning (DW-ATCL), and (iv) the Hate-Aware Semantic Mem-

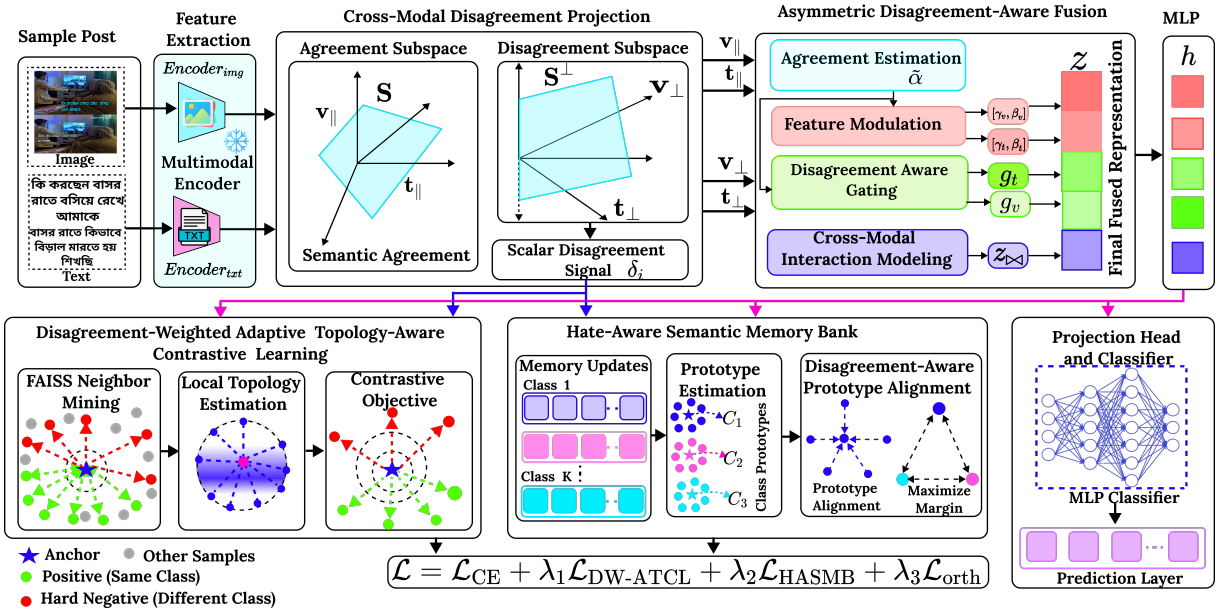


Figure 2: Overall architectural details of IndicHMNet.

ory Bank (HASMB). The detailed architecture is shown in Figure 2.

4.1 Multimodal Backbone

We use SigLIP2 as a frozen multimodal encoder to obtain image and text representations. Given an image–text pair (I, T), the encoder produces:

$$\boldsymbol{t}, \boldsymbol{v} = \text{SigLIP2}(I, T), \qquad \boldsymbol{t}, \boldsymbol{v} \in \mathbb{R}^{d_e}, \quad (1)$$

where $\boldsymbol{t}$ and $\boldsymbol{v}$ denote normalized text and image embeddings, respectively.

4.2 Cross-Modal Disagreement Projection

Vision–language embedding spaces often exhibit modality gaps and anisotropic geometry (Liang et al., 2022; Ethayarajh, 2019), making direct cross-modal alignment unreliable. CMDP addresses this issue by decomposing multimodal representations into shared and modality-specific subspaces.

Given normalized embeddings $\boldsymbol{t}, \boldsymbol{v} \in \mathbb{R}^{d_e}$, CMDP projects them into a latent space of dimension d_s using a shared projector W_a and modality-specific projectors $W_t^{\perp}$ and $W_v^{\perp}$:

$$\boldsymbol{t}_{\parallel} = \text{LN}(W_a \boldsymbol{t}), \qquad \boldsymbol{t}_{\perp} = \text{LN}(W_t^{\perp} \boldsymbol{t}), \quad (2)$$

$$\boldsymbol{v}_{\parallel} = \text{LN}(W_a \boldsymbol{v}), \qquad \boldsymbol{v}_{\perp} = \text{LN}(W_v^{\perp} \boldsymbol{v}), \quad (3)$$

where $(\boldsymbol{t}_{\parallel}, \boldsymbol{v}_{\parallel})$ denote shared representations and $(\boldsymbol{t}_{\perp}, \boldsymbol{v}_{\perp})$ capture modality-specific residual semantics. Since both modalities share the projector W_a, the disagreement components are learned relative to a common latent space, enabling CMDP to capture modality-specific deviations from shared multimodal semantics. Cross-modal disagreement is estimated from the disagreement representations:

$$\delta_i = \tanh\left(\frac{1}{2}\left(\frac{\|\boldsymbol{t}_{\perp,i}\|_2}{\sqrt{d_s}} + \frac{\|\boldsymbol{v}_{\perp,i}\|_2}{\sqrt{d_s}}\right)\right), \quad (4)$$

where larger δ_i values indicate stronger semantic divergence between modalities. We refer to δ_i as the scalar disagreement signal.

To encourage subspace disentanglement, CMDP applies the following covariance decorrelation objective:

$$\mathcal{L}_{\text{orth}} = \|\text{Cov}(\boldsymbol{t}_{\parallel}, \boldsymbol{t}_{\perp})\|_F^2 + \|\text{Cov}(\boldsymbol{v}_{\parallel}, \boldsymbol{v}_{\perp})\|_F^2 + \|\text{Cov}(\boldsymbol{t}_{\parallel}, \boldsymbol{v}_{\perp})\|_F^2 + \|\text{Cov}(\boldsymbol{v}_{\parallel}, \boldsymbol{t}_{\perp})\|_F^2. \quad (5)$$

where $\text{Cov}(\cdot, \cdot)$ denotes empirical cross-covariance estimated over a mini-batch. The resulting disagreement score is later used for adaptive contrastive weighting and prototype alignment.

4.3 Asymmetric Disagreement-Aware Fusion

The proposed Asymmetric Disagreement-Aware Fusion (ADAF) module operates on the shared and modality-specific representations generated by CMDP. Instead of relying on symmetric aggregation, ADAF conditions fusion using agreement-aware modulation, adaptive gating, and multiplicative cross-modal interactions.

Agreement-aware FiLM modulation: Since $\boldsymbol{t}_{\parallel}$

and $\boldsymbol{v}_{\parallel}$ lie in the shared latent space, their cosine similarity estimates semantic agreement:

$$\alpha = \cos(\boldsymbol{t}_{\parallel}, \boldsymbol{v}_{\parallel}), \qquad \tilde{\alpha} = \sigma(\alpha), \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid activation. The agreement score conditions both modalities through FiLM modulation (Perez et al., 2018):

$$[\boldsymbol{\gamma}_t, \boldsymbol{\beta}_t] = \phi_t(\tilde{\alpha}), \qquad [\boldsymbol{\gamma}_v, \boldsymbol{\beta}_v] = \phi_v(\tilde{\alpha}), \tag{7}$$

where $\phi_t(\cdot)$ and $\phi_v(\cdot)$ generate modulation parameters for both modalities.

Disagreement-aware gating: ADAF constructs modality-specific contextual representations using all CMDP components:

$$\boldsymbol{c}_t = [\boldsymbol{t}_{\parallel}; \boldsymbol{v}_{\parallel}; \boldsymbol{t}_{\perp}; \boldsymbol{v}_{\perp}; \tilde{\alpha}], \tag{8}$$

$$\boldsymbol{c}_v = [\boldsymbol{v}_{\parallel}; \boldsymbol{t}_{\parallel}; \boldsymbol{v}_{\perp}; \boldsymbol{t}_{\perp}; \tilde{\alpha}]. \tag{9}$$

The fusion process is asymmetric because text and image streams use independently parameterized gating functions:

$$\boldsymbol{g}_t = \sigma(\mathrm{LN}(W_t \boldsymbol{c}_t)), \qquad \boldsymbol{g}_v = \sigma(\mathrm{LN}(W_v \boldsymbol{c}_v)), \tag{10}$$

where the gating vectors adaptively regulate modality contributions using both agreement and disagreement cues.

Cross-modal interaction modeling: To capture higher-order interactions, ADAF further introduces a multiplicative interaction stream:

$$\boldsymbol{z}_{\bowtie} = \mathrm{LN}\Big((U_t \boldsymbol{t}_{\parallel}) \odot (U_v \boldsymbol{v}_{\parallel})\Big), \tag{11}$$

where $\odot$ denotes element-wise multiplication.

Final fused representation: The final multimodal representation combines the gated modality streams with the interaction representation:

$$\begin{aligned} \boldsymbol{z} = \Big[& \boldsymbol{g}_t \odot \Big(\boldsymbol{\gamma}_t \odot (\boldsymbol{t}_{\parallel} + \boldsymbol{t}_{\perp}) + \boldsymbol{\beta}_t\Big); \\ & \boldsymbol{g}_v \odot \Big(\boldsymbol{\gamma}_v \odot (\boldsymbol{v}_{\parallel} + \boldsymbol{v}_{\perp}) + \boldsymbol{\beta}_v\Big); \boldsymbol{z}_{\bowtie} \Big]. \end{aligned} \tag{12}$$

The fused representation preserves shared semantics, modality-specific cues, and cross-modal interactions for hateful meme classification.

4.4 Projection Head and Classifier

The fused representation $\boldsymbol{z} \in \mathbb{R}^{3d_s}$ is projected through a two-layer MLP:

$$\boldsymbol{h} = \mathrm{MLP}(\boldsymbol{z}), \qquad \boldsymbol{h} \in \mathbb{R}^D, \quad D = 512, \tag{13}$$

where $\boldsymbol{h}$ denotes the shared embedding used for downstream optimization objectives.

A linear classifier produces prediction logits:

$$\hat{\boldsymbol{y}} = W_c \boldsymbol{h} + \boldsymbol{b}_c, \tag{14}$$

where W_c and $\boldsymbol{b}_c$ are classifier parameters.

The classifier is optimized using label-smoothed cross-entropy:

$$\mathcal{L}_{\mathrm{CE}} = -\sum_{c=1}^{C} y_c^{\mathrm{LS}} \log(\hat{y}_c), \tag{15}$$

where y_c^{LS} denotes the smoothed target distribution with a smoothing factor 0.05.

4.5 Disagreement-Weighted Adaptive Topology-aware Contrastive Learning

DW-ATCL extends supervised contrastive learning using (i) local semantic topology estimation, (ii) class-adaptive temperature scaling, and (iii) disagreement-aware sample weighting.

Local semantic topology estimation: Given normalized embeddings $\boldsymbol{h}_i \in \mathbb{R}^D$, DW-ATCL estimates local semantic density using nearest-neighbor similarity structure. For each anchor $\boldsymbol{h}_i$, the local density is computed as:

$$\rho_i = \frac{1}{k} \sum_{j \in \mathcal{N}_k(i)} \boldsymbol{h}_i^{\top} \boldsymbol{h}_j, \tag{16}$$

where $\mathcal{N}_k(i)$ denotes the set of k nearest neighbors.

Class-adaptive temperature scaling: Instead of using a fixed global temperature, DW-ATCL learns class-specific temperature parameters:

$$\tau_0(c) = \exp(\theta_c^{\tau}), \qquad \beta(c) = \exp(\theta_c^{\beta}), \tag{17}$$

where θ_c^{τ} and θ_c^{β} are learnable class-wise parameters. The adaptive temperature for anchor i is:

$$\tau_i = \mathrm{clip}\Big(\tau_0(y_i) \exp(-\beta(y_i)\rho_i),\, 0.05,\, 0.5\Big), \tag{18}$$

where $\mathrm{clip}(\cdot)$ constrains the temperature range for stable optimization. This formulation sharpens contrastive discrimination in dense semantic neighborhoods while avoiding unstable optimization in sparse regions.

Disagreement-aware weighting: To emphasize semantically divergent memes, each sample receives a disagreement-aware weighting coefficient: $w_i = 0.5 + \delta_i$,where larger disagreement values receive stronger contrastive weighting.

Contrastive objective: Given a learnable global logit scaling factor s, the final topology-aware contrastive objective is:

$$\mathcal{L}_{\text{DW-ATCL}} = \frac{1}{B'} \sum_{i=1}^{B'} -w_i \log(p_i), \tag{19}$$

$$p_i = \frac{\exp\left(s\, \boldsymbol{h}_i^\top \boldsymbol{h}_i^+ / \tau_i\right)}{\exp\left(s\, \boldsymbol{h}_i^\top \boldsymbol{h}_i^+ / \tau_i\right) + \sum_{n=1}^{N_{\text{neg}}} \exp\left(s\, \boldsymbol{h}_i^\top \boldsymbol{h}_{i,n}^- / \tau_i\right)}. \tag{20}$$

Here, $\boldsymbol{h}_i^+$ denotes a positive neighbor from the same class, while $\{\boldsymbol{h}_{i,n}^-\}_{n=1}^{N_{\text{neg}}}$ are hard negatives retrieved from different classes. The resulting objective encourages topology-aware representation learning by combining local semantic density, adaptive temperature scaling, and disagreement-aware weighting.

4.6 Hate-Aware Semantic Memory Bank

While disagreement-aware contrastive learning improves local representation structure, hateful meme understanding also benefits from stable class-level semantic anchors. To encourage globally coherent representations, we introduce the Hate-Aware Semantic Memory Bank (HASMB), which maintains class-specific prototype memories for disagreement-aware prototype alignment.

Class-wise memory bank: HASMB maintains a fixed-capacity memory bank for each class: $\mathcal{M} = \{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_C\}$, where $\mathcal{M}_c = \{\boldsymbol{m}_{c,1}, \boldsymbol{m}_{c,2}, \ldots, \boldsymbol{m}_{c,K}\} \subset \mathbb{R}^D$. and C denotes the number of classes, K the memory capacity per class, and D the projection dimension. After each optimization step, normalized embeddings are inserted into the corresponding class memory queue.

Prototype estimation: The semantic prototype for class c is computed as the normalized mean representation of its memory bank:

$$\boldsymbol{p}_c = \frac{\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{m}_{c,k}}{\left\| \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{m}_{c,k} \right\|_2}. \tag{21}$$

The resulting prototypes provide stable class-level semantic anchors across training iterations.

Disagreement-aware prototype alignment: Given a projected embedding $\boldsymbol{h}_i$ with class label y_i, HASMB encourages alignment with the corresponding class prototype while emphasizing disagreement-heavy samples through CMDP disagreement weighting:

$$\mathcal{L}_{\text{pull}} = \frac{1}{B} \sum_{i=1}^{B} (0.5 + \delta_i) \left\| \hat{\boldsymbol{h}}_i - \boldsymbol{p}_{y_i} \right\|_2^2, \tag{22}$$

where

$$\hat{\boldsymbol{h}}_i = \frac{\boldsymbol{h}_i}{\|\boldsymbol{h}_i\|_2}$$

denotes the normalized embedding representation.

Inter-prototype separation: To encourage semantic separability between hateful meme categories, HASMB additionally penalizes similarity between class prototypes:

$$\mathcal{L}_{\text{push}} = \frac{1}{C(C-1)} \sum_{i \neq j} \cos(\boldsymbol{p}_i, \boldsymbol{p}_j). \tag{23}$$

The final prototype alignment objective is:

$$\mathcal{L}_{\text{HASMB}} = \mathcal{L}_{\text{pull}} + \lambda_{\text{sep}} \mathcal{L}_{\text{push}}, \tag{24}$$

where λ_{sep} controls inter-class prototype separation.

Overall training objective: The final optimization objective combines classification, disagreement-aware contrastive learning, prototype alignment, and subspace decorrelation:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda_1 \mathcal{L}_{\text{DW-ATCL}} + \lambda_2 \mathcal{L}_{\text{HASMB}} + \lambda_3 \mathcal{L}_{\text{orth}}. \tag{25}$$

We set $\lambda_1 = \lambda_2 = 0.01$ and $\lambda_3 = 0.05$ in all experiments.

5 Experimental Results

5.1 Performance Evaluation

We compare IndicHMNet against multilingual unimodal models, prompt-based vision–language models, and recent multimodal hateful meme detection frameworks. Performance is evaluated using Accuracy (Acc), Macro Precision (m-Prec), Macro Recall (m-Rec), and Macro-F1 (m-F1).

Task 1: Hate vs Non-Hate Classification: Table 4 presents results for binary hateful meme detection. Multilingual text models outperform vision-only baselines, highlighting the importance of textual cues in multilingual memes. However, unimodal methods remain limited because hateful intent often emerges from cross-modal interaction. Prompt-based vision–language models also perform poorly in zero-shot and few-shot

Method	Task 1				Task 2			
	Acc	m-Prec	m-Rec	m-F1	Acc	m-Prec	m-Rec	m-F1
Unimodal Models								
mBERT (Pires et al., 2019)	76.08	76.19	76.08	76.06	61.08	62.69	59.95	60.83
MuRIL (Khanuja et al., 2021)	77.94	78.15	77.95	77.91	62.94	65.66	59.70	60.93
XLM-R (Conneau et al., 2020)	77.48	77.72	77.48	77.43	60.42	59.74	58.05	58.15
ViT (Dosovitskiy et al., 2021)	70.68	71.35	70.69	70.45	56.71	54.55	55.23	55.11
ConvNeXt (Liu et al., 2022)	70.42	70.62	70.42	70.35	56.57	54.92	54.68	54.56
Vision–Language Models (Prompt-based)								
Gemma-3-4B-IT (0-shot) (Team et al., 2025)	60.40	66.50	60.40	56.37	40.00	48.83	40.00	35.11
Gemma-3-4B-IT (3-shot) (Team et al., 2025)	63.40	63.67	63.40	63.22	45.20	58.87	45.20	39.69
Qwen2.5-VL-7B-Instruct (0-shot) (Qwen et al., 2025)	63.40	63.58	63.40	63.28	38.80	53.25	38.80	36.25
Qwen2.5-VL-7B-Instruct (3-shot) (Qwen et al., 2025)	65.00	72.29	65.00	61.88	46.80	49.15	46.80	42.67
LLaMA-3.2-11B-Vision-Instruct (0-shot) (Grattafiori et al., 2024)	54.60	66.78	54.60	44.53	28.80	43.93	28.80	25.05
LLaMA-3.2-11B-Vision-Instruct (3-shot) (Grattafiori et al., 2024)	58.00	55.80	55.20	54.90	34.80	43.45	34.80	28.74
Multimodal Models								
SigLIP (Zhai et al., 2023)	85.21	85.22	85.31	85.21	73.30	72.54	72.88	72.67
RGCL (Mei et al., 2024)	73.75	74.33	73.76	73.60	62.28	61.30	59.59	60.04
MemeCLIP (Shah et al., 2024)	84.08	84.30	84.08	84.05	74.94	74.37	74.09	74.19
MHSDF (Prabhu et al., 2025)	83.74	83.74	83.75	83.74	70.65	69.92	71.08	70.36
TM-MHS (Kapil et al., 2025)	85.07	85.22	85.08	85.06	73.76	73.32	72.97	73.10
RA-HMD (Mei et al., 2025)	78.08	78.08	78.08	78.08	65.07	64.08	65.79	64.74
IndicHMNet	**87.87**	**87.87**	**87.87**	**87.87**	**77.82**	**77.68**	**76.77**	**77.09**

Table 4: Performance comparison with state-of-the-art methods on IndicHM. Best results are shown in **bold**.

Method	MUTE Dataset				BHM Dataset			
	Acc	m-Prec	m-Rec	m-F1	Acc	m-Prec	m-Rec	m-F1
Unimodal Models								
mBERT (Pires et al., 2019)	61.77	58.44	57.55	57.57	62.58	58.36	56.65	56.33
MuRIL Khanuja et al., 2021	67.54	66.72	60.78	60.17	67.09	64.34	61.99	62.23
XLM-R (Conneau et al., 2020)	65.14	63.19	57.64	56.08	67.83	61.41	57.54	56.65
ViT (Dosovitskiy et al., 2021)	67.54	65.30	63.30	63.60	66.24	63.72	63.36	63.50
ConvNeXt (Liu et al., 2022)	66.34	64.36	64.36	64.36	63.85	60.79	60.16	60.32
Vision–Language Models (Prompt-based)								
Gemma-3-4B-IT (0-shot) (Team et al., 2025)	59.25	58.80	59.25	58.99	60.90	61.45	60.90	61.14
Gemma-3-4B-IT (3-shot) (Team et al., 2025)	63.05	62.60	63.05	62.78	61.92	62.21	61.92	62.05
Qwen2.5-VL-7B-Instruct (0-shot) (Qwen et al., 2025)	59.04	57.45	57.61	57.49	58.65	56.03	56.08	56.05
Qwen2.5-VL-7B-Instruct (3-shot) (Qwen et al., 2025)	63.95	61.88	61.24	61.38	62.87	60.80	60.82	60.81
LLaMA-3.2-11B-Vision-Instruct (0-shot) (Grattafiori et al., 2024)	40.96	46.41	47.72	38.34	44.87	50.24	50.21	44.46
LLaMA-3.2-11B-Vision-Instruct (3-shot) (Grattafiori et al., 2024)	46.21	49.46	48.66	43.77	47.23	51.51	52.09	47.71
Multimodal Models								
SigLIP (Zhai et al., 2023)	63.46	63.87	64.67	63.09	64.42	61.56	61.06	61.22
RGCL (Mei et al., 2024)	68.99	67.33	64.11	64.43	66.67	64.00	60.29	60.12
MemeCLIP (Shah et al., 2024)	71.36	69.69	68.83	69.15	69.18	67.25	64.37	64.86
MHSDF (Prabhu et al., 2025)	66.35	64.73	64.53	64.62	67.48	63.03	65.25	63.48
TM-MHS (Kapil et al., 2025)	64.51	64.18	60.26	61.05	67.60	67.53	67.20	67.34
RA-HMD (Mei et al., 2025)	67.31	65.77	66.23	65.92	65.68	62.53	60.41	60.52
IndicHMNet	**73.08**	**72.52**	**73.77**	**72.50**	**73.12**	**71.68**	**72.45**	**71.93**

Table 5: Performance comparison with state-of-the-art methods on the MUTE and BHM datasets.

Ablation	Task 1: Binary-class				Task 2: Multi-class			
	Acc	m-Prec	m-Rec	m-F1	Acc	m-Prec	m-Rec	m-F1
Full Model (All Components)	**87.87**	**87.87**	**87.87**	**87.87**	**77.82**	**77.68**	**76.77**	**77.09**
w/o CMDP	85.55	85.54	85.55	85.55	75.43	75.24	74.97	75.06
w/o DW-ATCL	86.34	86.35	86.34	86.34	77.16	77.16	76.49	76.75
w/o HASMB	86.41	86.43	86.41	86.41	74.63	74.88	74.63	74.65
w/o Disagreement-aware Weighting	86.48	86.48	86.48	86.48	76.23	75.81	75.30	75.52
w/o Orthogonality	86.88	86.96	86.88	86.87	76.49	76.01	76.40	76.19
Static Tau (Fixed Temperature)	86.94	87.05	86.94	86.93	76.42	75.81	76.48	76.09
w/o ADAF (Simple Concat)	86.14	86.13	86.14	86.14	75.56	75.26	74.94	74.95

Table 6: Ablation study results for Task 1 and Task 2.

settings, indicating difficulty in handling culturally grounded and code-mixed hateful content without task-specific adaptation. Multimodal methods consistently outperform unimodal baselines, confirming the importance of joint image–text reasoning. IndicHMNet achieves the best overall performance, improving Macro-F1 by 2.81% over TM-MHS and 4.13% over MHSDF. The gains mainly stem from explicitly modeling shared agreement and modality-specific residual seman-

tics. CMDP preserves informative residual cues, ADAF improves disagreement-aware fusion, while DW-ATCL and HASMB produce discriminative and semantically structured representations.

Task 2: Hate Target Classification: Task 2 is more challenging because the model must identify fine-grained hate targets rather than only detect hateful intent. Consequently, performance decreases across all methods compared to Task 1. IndicHMNet achieves the best performance across all metrics, improving Macro-F1 by 3.99% over TM-MHS and 6.73% over MHSDF. These improvements indicate that disagreement-aware fusion and representation learning better preserve subtle target-specific semantics that are often weakened by conventional multimodal aggregation.

5.2 Evaluation on External Benchmarks

Table 5 reports results on the MUTE and BHM datasets. IndicHMNet consistently outperforms unimodal, prompt-based, and multimodal baselines across both benchmarks. On MUTE, IndicHMNet improves Macro-F1 by 3.35% over MemeCLIP and 6.58% over RA-HMD. On BHM, it improves Macro-F1 by 4.59% over TM-MHS and 7.07% over MemeCLIP. These results demonstrate that the proposed disagreement-centric framework generalizes effectively across datasets with different linguistic and multimodal characteristics.

5.3 Ablation Study

Table 6 reports ablation results for Task 1 and Task 2. Removing CMDP causes the largest performance drop, highlighting the importance of disagreement-aware representation decomposition. Replacing ADAF with simple concatenation also reduces performance, confirming the effectiveness of disagreement-aware fusion. Similarly, removing DW-ATCL, disagreement weighting, adaptive temperature scaling, orthogonality regularization, or HASMB consistently degrades performance, particularly for fine-grained hate-target classification.

6 Conclusion

We introduce **IndicHM**, a large-scale trilingual hateful meme benchmark containing over 15K manually annotated image–text pairs across English, Hindi, and Bengali for hateful intent and target classification. We further propose **IndicHMNet**, a disagreement-centric framework that models residual cross-modal semantics through shared agreement and modality-specific residual subspaces. The framework integrates disagreement-aware fusion, topology-aware contrastive learning, and prototype alignment within a unified objective. Experiments across languages and tasks show consistent improvements over strong multimodal and vision–language baselines, highlighting the importance of modeling modality-specific semantic divergence for multilingual meme understanding.

Limitations

Our work has several limitations. First, IndicHM covers only English, Hindi, and Bengali, leaving other widely used Indic languages for future extension. Second, each meme is assigned a single dominant hate target, although real-world content may involve multiple overlapping targets. Third, the disagreement score is estimated through residual semantic projections rather than explicit semantic reasoning. Finally, the framework does not generate explicit explanations, limiting interpretability for culturally implicit or symbolic hateful content.

Ethical Considerations

Data Collection and Privacy

IndicHM contains memes collected from publicly accessible sources such as Reddit and Pinterest, following applicable platform policies. It includes only meme images and embedded text and excludes personally identifiable information. No attempt was made to identify or track individuals.

Bias and Annotation

Potential biases may arise from cultural context, annotation subjectivity, and source selection. To reduce these, multilingual annotators proficient in English, Hindi, and Bengali followed standardized guidelines with multiple review rounds. Broad keywords were used for collection; however, residual biases may remain.

Intended Use

IndicHM is intended for academic, non-commercial research on multimodal hate detection and multilingual learning, particularly for low-resource Indic languages. It is not intended for profiling, surveillance, or harmful applications, and real-world use should include human oversight.

References

Hedi Ben-younes, Remi Cadene, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *2017 IEEE International Conference on Computer Vision (ICCV)*, page 2631–2639. IEEE.

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi3Hate: Multimodal, Multilingual, and Multicultural Hate Speech Detection with Vision–Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 9714–9731. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

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

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

Kawin Ethayarajh. 2019. How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, page 55–65. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Weikuo Guo, Huaibo Huang, Xiangwei Kong, and Ran He. 2019. Learning Disentangled Representation for Cross-Modal Retrieval with Deep Mutual Information Estimation. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, page 1712–1720, New York, NY, USA. Association for Computing Machinery.

Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1122–1131, New York, NY, USA. Association for Computing Machinery.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, Los Alamitos, CA, USA. IEEE Computer Society.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the Underlying Meaning of Multimodal Hateful Memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5995–6003. International Joint Conferences on Artificial Intelligence Organization.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A Multimodal Dataset for Detecting Hateful Memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, page 32–39. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah M. Preum. 2024. Deciphering Hate: Identifying Hateful Memes and Their Targets. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8347–8359, Bangkok, Thailand. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Prashant Kapil and Asif Ekbal. 2025. A transformer based multi task learning approach to multimodal hate speech detection. *Natural Language Processing Journal*, 11:100133.

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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, pages 18661–18673, Red Hook, NY, USA. Curran Associates Inc.

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, pages 2611–2624, Red Hook, NY, USA. Curran Associates Inc.

Gitanjali Kumari, Dibyanayan Bandyopadhyay, Asif Ekbal, and Vinutha B. NarayanaMurthy. 2024. CM-Off-Meme: Code-Mixed Hindi-English Offensive Meme Detection with Multi-Task Learning by Leveraging Contextual Knowledge. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3380–3393, Turin, Italy. European Language Resources Association (ELRA) and ICCL.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. *Preprint*, arXiv:1908.03557.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the Gap: Understanding the Modality Gap in Multi-Modal Contrastive Representation Learning. In *Advances in Neural Information Processing Systems 35*, NeurIPS 2022, page 17612–17625. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Advances in neural information processing systems*, 32.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

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.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. 2025. Robust Adaptation of Large Multimodal Models for Retrieval Augmented Hateful Meme Detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23806–23828, Suzhou, China. Association for Computational Linguistics.

Shreyash Mishra, S Suryavardan, Parth Patwa, Megha Chakraborty, Anku Rani, Aishwarya Reganti, Aman Chadha, Amitava Das, Amit Sheth, Manoj Chinnakotla, Asif Ekbal, and Srijan Kumar. 2023. Memotion 3: Dataset on Sentiment and Emotion Analysis of Codemixed Hindi-English Memes. *Preprint*, arXiv:2303.09892.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 Task 5: Toxic Spans Detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. 2018. FiLM: Visual Reasoning with a General Conditioning Layer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How Multilingual is Multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

R. Prabhu and V. Seethalakshmi. 2025. A comprehensive framework for multi-modal hate speech detection in social media using deep learning. *Scientific Reports*, 15(1).

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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 *International conference on machine learning*, pages 8748–8763. PmLR.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. MemeCLIP: Leveraging CLIP Representations for Multimodal Meme Classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332, Miami,

Florida, USA. Association for Computational Linguistics.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. MIMIC: Misogyny Identification in Multimodal Internet Content in Hindi-English Code-Mixed Language. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 Technical Report. *Preprint*, arXiv:2503.19786.

Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. 2021. Learning Modality-Specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(12):10790–10797.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952.

A Additional Related Work

We further discuss prior work on multimodal representation learning, modality gap mitigation, and contrastive multimodal reasoning, which form the technical foundation of the proposed disagreement-centric framework.

A.1 Multimodal Hate Detection Architectures

Existing multimodal hate detection methods primarily focus on improving cross-modal fusion and multimodal representation learning. Early approaches adopted transformer-based vision–language architectures such as VisualBERT, ViLBERT, and OSCAR for joint image–text reasoning. More recent methods including MemeCLIP (Shah et al., 2024), RGCL (Mei et al., 2024), RA-HMD (Mei et al., 2025), TM-MHS (Kapil and Ekbal, 2025), and MHSDF (Prabhu and Seethalakshmi, 2025) incorporate contrastive learning, retrieval augmentation, semantic disentanglement, and target-aware fusion for hateful meme understanding. Parallel work in hate speech detection has also explored explainability and fine-grained reasoning through benchmarks such as HateXplain (Mathew et al., 2021), ToxicSpans (Pavlopoulos et al., 2021).

A.2 Modality gap and representation anisotropy

Vision–language representations often exhibit a persistent modality gap (Liang et al., 2022) and geometric anisotropy (Ethayarajh, 2019), where embeddings from different modalities occupy structurally distinct regions of the latent space. Consequently, direct embedding comparisons may conflate semantic inconsistency with modality-specific variation. Several works have explored reducing this gap through contrastive alignment and shared latent representation learning. CLIP (Radford et al., 2021) demonstrates the effectiveness of large-scale contrastive vision–language pretraining, while ALIGN (Jia et al., 2021) further improves multimodal alignment using large-scale noisy image–text pairs. Other studies investigate disentangled multimodal representations by separating shared and modality-specific information. Several studies have explored disentangled multimodal representations by separating shared and modality-specific information for cross-modal learning and multimodal sentiment analysis (Guo et al., 2019; Yu et al., 2021; Hazarika et al., 2020). In contrast, the proposed CMDP explicitly models modality disagreement as a meaningful semantic signal for hateful meme understanding rather than solely minimizing modality discrepancy.

A.3 Multimodal fusion and interaction modeling

A large body of multimodal learning research focuses on improving cross-modal fusion and interaction modeling. Early transformer-based approaches such as VisualBERT (Li et al., 2019), ViLBERT (Lu et al., 2019), and LXMERT (Tan and Bansal, 2019) jointly encode visual and textual representations using cross-modal attention mechanisms. Subsequent methods explored feature modulation and bilinear interaction modeling to improve multimodal reasoning. FiLM (Perez et al., 2018) introduces feature-wise conditioning for adaptive representation modulation, while MUTAN (Ben-younes et al., 2017) employs bilinear Tucker decomposition for higher-order multimodal

interactions. Our ADAF module builds upon these ideas by combining agreement-aware modulation with disagreement-conditioned fusion for hateful meme reasoning.

A.4 Contrastive and topology-aware representation learning

Contrastive learning has become a dominant paradigm for multimodal representation learning and semantic alignment. Supervised Contrastive Learning (Khosla et al., 2020) demonstrates strong representation clustering through label-aware contrastive objectives, while SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) establish scalable contrastive representation learning frameworks. Recent multimodal hate detection methods also employ contrastive objectives to improve representation discrimination. Retrieval-guided contrastive learning methods such as RGCL (Mei et al., 2024) leverage semantic retrieval to improve multimodal hate detection, whereas other studies incorporate adaptive hard negative mining and semantic regularization. Our DW-ATCL extends this direction through disagreement-aware sample weighting and adaptive temperature scaling based on local semantic topology.

B Extended Dataset Statistics

We additionally evaluate our approach on two existing hateful meme datasets, BHM and MUTE, alongside the proposed IndicHM dataset. For BHM and MUTE, experiments are conducted only on Task 1, while IndicHM supports evaluation on both hate detection and target classification. BHM and MUTE contain approximately 7.5K and 4.1K Bengali hateful memes, respectively, including code-mixed variants.

B.1 Language and Source-wise Sample Distribution

Table 7 presents the distribution of samples across languages and data sources. The dataset is approximately balanced across Hindi, English, and Bengali, with nearly 5,000 samples per language. For each language, memes are collected from both Pinterest and Reddit to ensure diversity in linguistic style, meme composition, and platform-specific content characteristics.

B.2 Category-wise Sample Distribution

Table 8 presents the distribution of hateful and non-hateful memes across different data sources. The dataset is carefully balanced, with nearly equal numbers of hateful and non-hateful samples collected from both Pinterest and Reddit. This balanced distribution helps reduce category bias and enables more reliable evaluation across classification tasks.

Language	Pinterest	Reddit	Total
Hindi	2,867	2,142	5,009
English	2,860	2,135	4,995
Bengali	2,868	2,132	5,000
Total	8,595	6,409	15,004

Table 7: Language and source-wise distribution of samples in IndicHM.

Source	Hate	Non-Hate	Total
Pinterest	4,302	4,293	8,595
Reddit	3,202	3,207	6,409
Total	7,504	7,500	15,004

Table 8: Category-wise distribution of hateful and non-hateful samples across sources.

B.3 Dataset Splits

IndicHM supports two benchmark tasks: (i) hate vs non-hate classification, and (ii) hate target classification. For both tasks, the dataset is divided into training, validation, and test sets using an approximate 80:10:10 split ratio. Splits are constructed independently for each language to ensure balanced multilingual evaluation and to avoid cross-lingual leakage.

B.3.1 Task 1: Hate vs Non-Hate Classification

Task 1 focuses on binary classification, where each meme is categorized as either hateful or non-hateful. The task contains a total of 15,004 samples.

Table 9 presents the overall dataset split used for Task 1. The training split is used for model optimization, the validation split for hyperparameter tuning, and the test split for final evaluation.

Split	Samples	Percentage
Train	12,003	80.00%
Validation	1,500	10.00%
Test	1,501	10.00%

Table 9: Dataset split for Task 1 (Hate vs Non-Hate Classification).

Table 10 shows the language-wise distribution of samples for Task 1. Each language contributes approximately one-third of the total samples within every split, ensuring balanced multilingual training and evaluation.

Split	Hindi	English	Bengali
Train	4,007 (33.38%)	3,996 (33.29%)	4,000 (33.33%)
Validation	501 (33.40%)	499 (33.27%)	500 (33.33%)
Test	501 (33.38%)	500 (33.31%)	500 (33.31%)

Table 10: Language-wise distribution for Task 1.

B.3.2 Task 2: Hate Target Classification

Task 2 focuses on hate target classification, where hateful memes are categorized according to their primary target group. The task contains 7,504 hateful meme samples.

Table 11 presents the dataset split used for Task 2. Similar to Task 1, an approximate 80:10:10 partition ratio is maintained across training, validation, and test sets.

Split	Samples	Percentage
Train	6,001	79.97%
Validation	750	9.99%
Test	753	10.03%

Table 11: Dataset split for Task 2 (Hate Target Classification).

Table 12 presents the language-wise distribution for Task 2. The dataset maintains balanced representation across Hindi, English, and Bengali to support fair multilingual evaluation and reduce language-specific bias.

Split	Hindi	English	Bengali
Train	2,006 (33.43%)	1,995 (33.24%)	2,000 (33.33%)
Validation	251 (33.47%)	249 (33.20%)	250 (33.33%)
Test	252 (33.47%)	251 (33.33%)	250 (33.20%)

Table 12: Language-wise distribution for Task 2.

B.4 Code-Mixing and Transliteration Analysis

Table 13 summarizes the code-mixing characteristics of IndicHM across languages. A substantial portion of the Hindi and Bengali subsets contains code-mixed content, reflecting the multilingual and informal nature of South Asian social media communication. Hindi exhibits the highest degree of code-mixing, with 41.03% of memes containing mixed-language text, while Bengali contains 22.10% code-mixed samples. We further observe frequent script switching and transliteration, particularly in Hindi memes where Romanized Hindi is commonly mixed with English expressions. Bengali memes also exhibit notable script-switching behavior, although transliterated content is comparatively less frequent. Overall, the dataset contains 3,160 code-mixed memes, highlighting the linguistic diversity and noisy multimodal characteristics of real-world hateful meme discourse.

Lang	Total	CM		SS	TL
		Count	%		
English	4,995	–	–	–	–
Hindi	5,009	2,055	41.03	582	1,473
Bengali	5,000	1,105	22.10	1,046	59
Total	**15,004**	**3,160**	**21.06**	**1,628**	**1,532**

Table 13: Code-mixing statistics across languages in IndicHM. Here, CM: Code-Mixed, SS: Script-Switched, TL: Transliterated.

C Dataset Collection and Preprocessing

IndicHM was collected from publicly accessible social media platforms, including Reddit and Pinterest, where meme-based communication is highly prevalent. To minimize copyright-related concerns, we restricted data collection to publicly accessible pages and community discussion threads. The final corpus contains approximately 5,000 memes for each language subset: English, Hindi, and Bengali.

A major challenge during dataset construction arises from the noisy nature of multilingual meme text. Existing Optical Character Recognition (OCR) systems often fail on Indic scripts when text appears over complex visual backgrounds, stylized fonts, or low-resolution meme templates. This issue becomes substantially more severe in code-mixed and transliterated memes commonly observed in South Asian social media.

To address these limitations, we adopted a human-in-the-loop transcription pipeline instead of relying solely on automated OCR systems. Native speakers manually transcribed textual content from every meme while verifying consistency with the accompanying visual context. For transliterated Hindi and Bengali memes written in Roman script, annotators additionally standardized spelling variations and restored missing or fragmented tokens caused by image artifacts or partial occlusion. This

process enabled the construction of a cleaner and semantically reliable textual representation while still preserving the linguistic characteristics of real-world social media discourse.

D Annotation Guidelines

The annotation protocol was designed to capture both explicit and implicit forms of hateful expression commonly observed in multimodal meme communication. Since hateful intent in memes often emerges through the interaction between textual and visual cues, annotators were instructed to evaluate the meme holistically rather than relying on isolated keywords, symbols, or image content.

D.1 Task Definitions

Each meme was annotated for the following tasks:
Task 1: Hateful Intent Detection: A meme was labeled as Hateful if it conveyed hostility, demeaning rhetoric, exclusionary narratives, or incitement toward an individual or social group based on protected characteristics such as religion, gender, race/ethnicity, or political identity. Memes that did not express such intent were labeled as Non-Hateful.
Task 2: Target Classification: For memes identified as hateful, annotators further assigned the primary target category from the following classes: Religion, Gender, Race or Ethnicity, Politics, and Other.

D.2 Annotation Principles

Annotators were instructed to infer the *intended meaning jointly conveyed by the text and image*. In many cases, hateful intent was expressed indirectly through sarcasm, irony, coded language, symbolic imagery, or contextual associations rather than explicit slurs or abusive terms. Consequently, annotations were guided by pragmatic interpretation instead of surface-level lexical matching.

Political criticism, satire, or ideological disagreement was not automatically considered hateful. Such memes were labeled as hateful only when they targeted protected communities or promoted dehumanizing, exclusionary, or harmful narratives against identifiable groups. Memes that remained highly ambiguous or lacked sufficient contextual evidence to support reliable interpretation were excluded during dataset construction.

D.3 Annotation Procedure

The annotation team consisted of six native speakers, with two annotators assigned per language. All annotators were additionally fluent in English, enabling consistent interpretation of multilingual and code-mixed content.

Before full-scale annotation, annotators completed a calibration phase using a curated seed set of 200 memes containing difficult and borderline examples, including: implicit hate speech, sarcasm and irony, regional slang and coded expressions, symbolic political references, and transliterated code-mixed content. Disagreements arising during calibration were discussed in adjudication meetings supervised by two senior researchers. The annotation guidelines were iteratively refined based on these discussions to improve cross-lingual consistency and reduce subjective interpretation bias. During the main annotation phase, disagreements between annotators were resolved through discussion and final adjudication by a senior annotator.

D.4 Handling of Code-Mixed and Multilingual Content

A substantial portion of Hindi and Bengali memes contained English code-mixing and Romanized transliterations (e.g., Hinglish and Benglish). To improve annotation consistency, annotators normalized common phonetic spelling variations and resolved transliterated expressions while preserving the original semantic intent.

Annotators were additionally instructed to restore incomplete, stylized, or visually truncated words whenever contextual interpretation was sufficiently reliable. For memes where hateful intent was primarily conveyed through symbolic imagery or implicit cultural references rather than explicit textual content, annotators provided short contextual notes to support consistent labeling decisions across annotators.

E Compared Methods

We compare IndicHMNet against a broad set of unimodal, vision–language, and multimodal baselines to assess its effectiveness. For text-only models, we evaluate multilingual encoders, such as, mBERT, MuRIL, and XLM-R; for vision-only baselines, we use ViT and ConvNeXt. We further include large vision–language models such as Gemma-3-4B-IT, Qwen2.5-VL-7B-Instruct, and LLaMA-3.2-11B-Vision-Instruct. For supervised unimodal

and multimodal baselines, encoder features are fed to a lightweight feedforward classifier to ensure fair comparison. We also compare against recent multimodal hate detection methods, including SigLIP (Zhai et al., 2023), RGCL (Mei et al., 2024), MHSDF (Prabhu and Seethalakshmi, 2025), TM-MHS (Kapil and Ekbal, 2025), MemeCLIP (Shah et al., 2024) and RA-HMD (Mei et al., 2025). All models are evaluated under identical experimental settings for consistent comparison across datasets.

F Experimental Setup

Training Details: Models are trained for 50 epochs using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-2}. The batch size is set to 16. At the beginning of each epoch, a FAISS index is constructed over all normalized training embeddings. During training, the top-$k = 10$ nearest neighbors are retrieved for each anchor.

Implementation Details: All experiments were conducted on a server equipped with an Intel(R) Xeon(R) Silver 4316 CPU @ 2.30 GHz and four NVIDIA A16 GPUs, each with 16 GB of VRAM. The system was configured with CUDA 12.8 and NVIDIA driver version 570.133.07. Model training and evaluation were implemented using PyTorch with GPU acceleration.

The proposed IndicHMNet contains approximately 380.6M total parameters, of which 5.39M parameters are trainable and 375.2M belong to the frozen SigLIP2 backbone encoder. Among the trainable components, CMDP contains 1.18M parameters, ADAF contains 3.15M parameters, the projection MLP contains 1.05M parameters, and the classifier contains 1,026 parameters. Table 14 shows the hyperparameter details.

Hyperparameter	**Value**
Model Architecture	
Vision-Language Backbone	SigLIP2-Base
Embedding / Fused / Hidden Dim	768 / 1,536 / 512
Training	
Batch Size / Epochs	16 / 50
Learning Rate / Weight Decay	1×10^{-4} / 1×10^{-2}
Gradient Clipping / Label Smoothing	1.0 / 0.05
Dropout	0.1
ATCL	
Base Temperature (τ_{base}) / β	0.07 / 2.0
Neighbors (k) / Hard Negatives	10 / 5
HASMB	
Memory Bank Size / Momentum	1,024 / 0.99
λ_1 / λ_2 / λ_3	0.01 / 0.01 / 0.5
Optimizer	AdamW

Table 14: Hyperparameter settings for IndicHMNet framework.

G Qualitative and Failure Case Analysis

Figure 3 presents representative qualitative examples illustrating both correct predictions and failure cases for hate detection (Task 1) and target classification (Task 2). Overall, the model performs reliably when hateful intent is expressed through explicit visual or textual cues. For instance, memes containing overt slurs, symbolic imagery, or clearly aligned visual–textual signals are classified with high confidence.

Failure cases in Task 1 primarily arise in memes where hateful intent is conveyed implicitly through irony, sarcasm, or culturally specific references. In such examples, the image and text may appear benign in isolation, with hate emerging only from subtle cross-modal interaction. Although disagreement-aware fusion improves robustness in many such cases, the model occasionally underestimates implicit hostility, leading to false negatives. For Task 2, most errors occur between semantically related target categories, such as religion and politics, or gender and broader social stereotypes. These misclassifications reflect the inherent ambiguity of real-world memes, where a single meme may simultaneously reference multiple social identities or ideological contexts. Since our formulation assumes a single dominant target, such overlapping cases remain challenging.

Overall, the qualitative analysis highlights both the strengths and limitations of the proposed approach. While explicit modeling of cross-modal disagreement and retrieval-guided contrastive learning improves interpretability and robustness, accurately resolving implicit, multi-target, and culturally grounded hate remains an open challenge.

H Language-wise Analysis

Tables 15 and 17 present language-wise Task 1 hate detection results for pretrained unimodal baselines, while Table 16 reports IndicHMNet performance across both Task 1 and Task 2. The analysis reveals substantial variation across languages, highlighting the challenges of multilingual and culturally grounded hateful meme understanding.

English memes are comparatively more difficult for unimodal models, particularly for multilingual text encoders such as MuRIL, mBERT, and XLM-R. This suggests that English memes in IndicHM



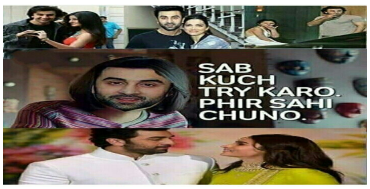
(a) Actual: Non-Hate
Pred: Non-Hate, Prob: 0.945



(b) Actual: Hate
Pred: Hate, Prob: 0.994



(c) Actual: Hate
Pred: Non-Hate, Prob: 0.752



(d) Actual: Gender
Pred: Gender, Prob: 0.821



(e) Actual: Racism
Pred: Racism, Prob: 0.996



(f) Actual: Religion
Pred: Political, Prob: 0.753

Figure 3: Qualitative examples illustrating model predictions for Task 1 (top row: hate detection) and Task 2 (bottom row: target classification). Each example reports the ground-truth label, predicted label, and model confidence. Errors primarily occur in cases involving implicit hate or overlapping target categories.

Model	English				Hindi			
	Acc	m-Prec	m-Rec	m-F1	Acc	m-Prec	m-Rec	m-F1
MuRIL	64.00	64.09	64.00	63.94	88.80	88.84	88.80	88.80
XLM-R	65.00	65.01	65.00	65.00	88.40	88.82	88.40	88.37
mBERT	64.20	64.24	64.20	64.18	88.20	88.23	88.20	88.20
ViT	65.80	65.81	65.80	65.80	90.20	90.21	90.20	90.20
ConvNeXt	64.60	64.60	64.60	64.60	90.00	90.13	90.00	89.99

Table 15: Analysis of English and Hindi Language on Task 1.

Language	Task 1				Task 2			
	Acc	m-Prec	m-Rec	m-F1	Acc	m-Prec	m-Rec	m-F1
English	82.28	81.26	81.25	81.25	86.82	86.55	82.56	85.15
Hindi	92.53	92.53	92.54	92.53	77.85	79.71	71.96	74.53
Bengali	92.08	92.07	92.08	92.08	74.95	74.86	74.95	74.72

Table 16: Language-wise performance on Task 1 and Task 2.

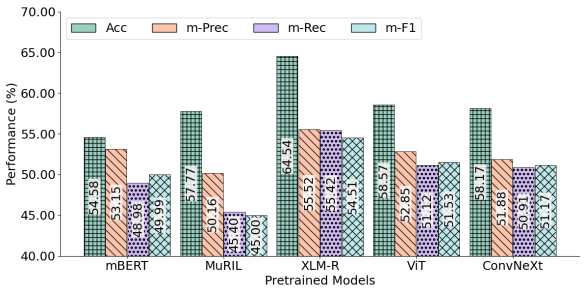


Figure 4: Analysis of English on Task 2

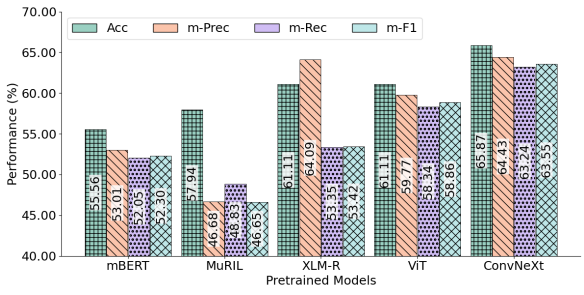


Figure 5: Analysis of Hindi on Task 2

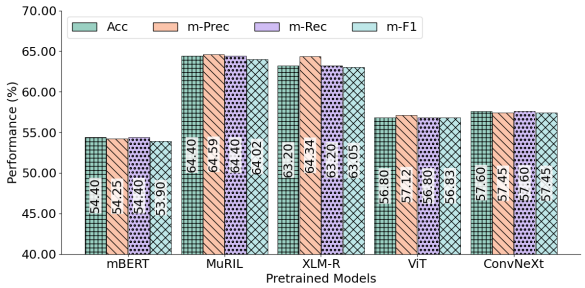


Figure 6: Analysis of Bengali on Task 2

Model	Bengali			
	Acc	m-Prec	m-Rec	m-F1
MuRIL	84.60	84.69	84.60	84.59
XLM-R	86.20	87.01	86.20	86.12
mBERT	88.20	88.42	88.20	88.18
ViT	87.80	88.12	87.80	87.77
ConvNeXt	90.40	90.42	90.44	90.28

Table 17: Analysis of Bengali Language on Task 1.

often rely on implicit semantics, sarcasm, and multimodal contextual cues that are difficult to capture from text alone. Unlike Hindi and Bengali, English memes frequently contain weaker lexical hate markers and stronger image–text incongruity, requiring deeper cross-modal contextual reasoning for correct interpretation. Vision-only encoders perform slightly better, indicating that visual semantics contribute more strongly in English hateful memes.

In contrast, Hindi and Bengali subsets achieve consistently higher Task 1 performance across both text and vision baselines, suggesting the presence of comparatively stronger lexical or visual hate indicators. Vision-based encoders, particularly ConvNeXt, achieve the strongest performance on Bengali memes, underscoring the importance of visual context in multilingual hateful meme understanding.

Table 16 further demonstrates that IndicHMNet generalizes effectively across all languages and tasks. Hindi and Bengali achieve the strongest Task 1 performance, whereas Task 2 remains considerably more challenging due to fine-grained hate target classification. The noticeable performance drop from Task 1 to Task 2 across all languages highlights the difficulty of modeling subtle target-specific semantics, especially in code-mixed and culturally contextual memes. Additional Task 2 language-wise trends are illustrated in Figures 4, 5, and 6.

Also, to ensure fair multilingual comparison, the same training configuration and hyperparameters were used across all languages, including learning rate, batch size, optimizer, and training schedule. This design isolates architectural effectiveness from language-specific hyperparameter tuning effects.

I Baseline Training Protocol

To ensure a fair comparison across different modalities and architectures, we adopt a consistent training protocol for all supervised baselines. For the unimodal language baselines, including mBERT, MuRIL, and XLM-R, the pretrained encoders are kept frozen, and only a lightweight MLP classifier is trained for the downstream classification task. Similarly, for the visual baselines, such as ViT, ConvNeXt, and SigLIP, the pretrained visual encoders remain frozen throughout training, while only the corresponding MLP classification head is optimized. Thus, the pretrained representations are kept fixed and the number of trainable parameters is restricted to the task-specific classifier.

For **IndicHMNet**, the pretrained SigLIP2 backbone is also kept frozen. Only the task-specific components, such as the projection layers, disagreement-aware fusion module, contrastive learning components, and classification heads, are optimized during training. This consistent protocol isolates the contribution of the proposed multimodal learning framework and avoids performance differences arising from unequal backbone fine-tuning.

J Diagnostic Evaluation and Annotation Limitations

J.1 Single-Dominant Target Annotation

Hateful memes can target multiple groups simultaneously, making target annotation inherently challenging. In the current IndicHM benchmark, each hateful meme is assigned a single dominant target to provide a well-defined and consistent target classification setting. Thus, the target annotation represents the primary target identified by the annotators rather than all potentially affected groups. This design is also consistent with several existing hateful meme benchmarks, such as Facebook Hateful Memes, MUTE, MIMIC, and BHM, which adopt single-label target annotations.

While this formulation provides a controlled evaluation setting, it does not fully capture the

multi-target nature of real-world hateful content. In particular, a meme may simultaneously involve related categories such as religion and politics or gender and social stereotypes. Extending IndicHM with multi-label target annotations would, therefore, provide a more realistic formulation of target classification and is an important direction for future work.

J.2 Diagnostic Evaluation across Linguistic Conditions

To further assess model robustness beyond the overall language-wise evaluation, we perform a fine-grained analysis under different linguistic conditions. Specifically, we divide the test samples into non-code-mixed and code-mixed subsets and separately evaluate script-switched and transliterated samples. This analysis is particularly relevant for Indic memes, where language mixing, script variation, and transliteration are common and can introduce additional challenges for multimodal hate detection.

Model	Hindi		Bengali	
	Non-CM	CM	Non-CM	CM
TM-MHS	83.48	77.74	94.95	93.98
SigLIP2	85.78	82.49	94.71	92.96
MemeCLIP	82.51	80.64	92.68	90.02
IndicHMNet	**86.11**	**84.74**	**96.71**	**95.03**

Table 18: Macro-F1 (%) comparison on non-code-mixed and code-mixed samples for Task 1. Here, CM = Code-Mixed, Non-CM = Non-Code-Mixed.

Task 1: Code-Mixed vs. Non-Code-Mixed Samples: Table 18 reports Macro-F1 scores for binary hate classification (Task 1) on code-mixed and non-code-mixed samples. IndicHMNet consistently achieves the highest performance across both Hindi and Bengali subsets. For Hindi, IndicHMNet obtains 86.11% and 84.74% Macro-F1 on non-code-mixed and code-mixed samples, respectively, while for Bengali it achieves 96.71% and 95.03%. Although code-mixing leads to a performance reduction for all models, IndicHMNet maintains a comparatively strong performance under mixed-language conditions.

Task 1: Script-Switched and Transliterated Samples: We further evaluate Task 1 on script-switched and transliterated samples, as reported in Table 19. IndicHMNet obtains the highest Macro-F1 on Hindi script-switched and transliterated samples, with scores of 86.94% and 84.09%, respectively. For Bengali, IndicHMNet achieves 96.89% on script-switched samples and 88.85% on transliterated samples. These results demonstrate that the proposed model remains effective when textual content deviates from its standard script representation.

Model	Hindi		Bengali	
	SS	Trans	SS	Trans
TM-MHS	76.55	74.00	95.77	80.00
SigLIP2	82.91	80.70	95.73	82.84
MemeCLIP	82.97	79.64	**98.96**	82.86
IndicHMNet	**86.94**	**84.09**	96.89	**88.85**

Table 19: Macro-F1 (%) comparison on script-switched and transliterated samples for Task 1 (binary hate classification). Here, SS = Script-Switched and Trans = Transliterated.

Task 2: Code-Mixed vs. Non-Code-Mixed Samples: We next analyze the more challenging hate target classification task (Task 2). As shown in Table 20, IndicHMNet achieves the highest Macro-F1 across both code-mixed and non-code-mixed subsets for Hindi and Bengali. IndicHMNet obtains 73.10% and 68.90% Macro-F1 on Hindi non-code-mixed and code-mixed samples, respectively, while achieving 74.54% and 67.49% on the corresponding Bengali subsets. The results indicate that the proposed model provides more consistent target classification under mixed-language conditions compared with the evaluated baselines.

Model	Hindi		Bengali	
	Non-CM	CM	Non-CM	CM
TM-MHS	71.48	62.03	72.55	63.21
SigLIP2	67.53	63.93	69.44	64.90
MemeCLIP	71.83	56.76	62.59	45.16
IndicHMNet	**73.10**	**68.90**	**74.54**	**67.49**

Table 20: Macro-F1 (%) comparison on non-code-mixed and code-mixed samples for Task 2 (hate target classification). Here, Non-CM = Non-Code-Mixed; CM = Code-Mixed.

Task 2: Script-Switched and Transliterated Samples: Finally, we evaluate target classification under script-switched and transliterated conditions. Table 21 shows that IndicHMNet consistently achieves the highest Macro-F1 across all four linguistic conditions. It obtains 64.24% and

64.57% on Hindi script-switched and transliterated samples, respectively, and 69.29% and 69.33% on the corresponding Bengali subsets. The consistent performance across these conditions suggests that combining visual and textual information helps mitigate the challenges introduced by script variation and transliteration.

Model	Hindi		Bengali	
	SS	Trans	SS	Trans
TM-MHS	60.64	45.67	64.20	60.52
SigLIP2	46.06	57.05	66.25	66.27
MemeCLIP	53.37	46.35	47.82	42.26
IndicHMNet	**64.24**	**64.57**	**69.29**	**69.33**

Table 21: Macro-F1 (%) comparison on script-switched and transliterated samples for Task 2. Here, SS = Script-Switched, Trans= Transliterated.

K Class-wise and Language-wise Performance Analysis

To provide a detailed evaluation of IndicHMNet, we analyze its class-wise performance for both tasks and further report language-wise results for English, Hindi, and Bengali.

K.1 Overall Class-wise Performance

For Task 1, which performs binary hate detection, IndicHMNet achieves an overall accuracy and Macro-F1 of 87.87%. The model obtains an F1-score of 87.87% for the Non-Hate class and 87.88% for the Hate class, indicating highly balanced performance between the two classes. The corresponding precision and recall values are also closely matched, as shown in Table 22.

Class	Prec	Rec	F1
Non-Hate	87.87	87.87	87.87
Hate	87.88	87.88	87.88
Macro Avg.	87.87	87.87	87.87
Acc			87.87

Table 22: Class-wise performance (%) of IndicHMNet on Task 1.

For Task 2, which classifies hateful memes into five target categories, IndicHMNet achieves an accuracy of 77.82% and a Macro-F1 of 77.09%. The highest F1-score is obtained for the Racism category (83.02%), followed by Gender Based (82.01%), Religion (79.01%), Political (75.68%), and Others (65.75%). The relatively lower performance on the Others category suggests that this class is more difficult to distinguish from the remaining target categories. The results are shown in Table 23.

Class	Prec	Rec	F1
Gender Based	78.44	85.93	82.01
Others	72.00	60.50	65.75
Political	75.17	76.19	75.68
Racism	83.81	82.24	83.02
Religion	79.01	79.01	79.01
Macro Avg.	77.68	76.77	77.09
Acc			77.82

Table 23: Class-wise performance (%) of IndicHMNet on Task 2.

K.2 Language-wise Performance:

We further evaluate IndicHMNet separately on English, Hindi, and Bengali samples to examine whether model performance varies across languages. For Task 1, Bengali achieves the highest performance, with an accuracy and Macro-F1 of 96.40%. Hindi follows with 87.23% accuracy and 87.22% Macro-F1, while English obtains 80.00% accuracy and 79.98% Macro-F1. The results are shown in Table 24.

Language	Class	Prec	Rec	F1
English	Non-Hateful	81.78	77.20	79.42
	Hateful	78.41	82.80	80.54
	Macro Avg.	80.09	80.00	79.98
	Acc			80.00
Hindi	Non-Hateful	85.77	89.20	87.45
	Hateful	88.80	85.26	86.99
	Macro Avg.	87.28	87.23	87.22
	Acc			87.23
Bengali	Non-Hateful	95.67	97.20	96.43
	Hateful	97.15	95.60	96.37
	Macro Avg.	96.41	96.40	96.40
	Acc			96.40

Table 24: Language-wise class-wise performance (%) of IndicHMNet on Task 1.

For Task 2, English obtains the highest language-wise performance with 85.66% accuracy and 81.20% Macro-F1. Hindi achieves 76.19% accuracy and 75.51% Macro-F1, whereas Bengali obtains 71.60% accuracy and 71.20% Macro-F1. The results indicate that target classification is more challenging than binary hate detection, with greater

Language	Class	Prec	Rec	F1
English	Gender Based	83.15	91.36	87.06
	Others	66.67	48.00	55.81
	Political	90.00	87.80	88.89
	Racism	81.82	79.41	80.60
	Religion	92.96	94.29	93.62
	Macro Avg.	82.92	80.17	81.20
	Acc			85.66
Hindi	Gender Based	77.33	85.29	81.12
	Others	80.00	63.64	70.89
	Political	72.22	69.64	70.91
	Racism	84.21	69.57	76.19
	Religion	73.91	83.61	78.46
	Macro Avg.	77.54	74.35	75.51
	Acc			76.19
Bengali	Gender Based	72.22	78.00	75.00
	Others	68.09	64.00	65.98
	Political	67.27	74.00	70.48
	Racism	84.91	90.00	87.38
	Religion	63.41	52.00	57.14
	Macro Avg.	71.18	71.60	71.20
	Acc			71.60

Table 25: Language-wise class-wise performance (%) of IndicHMNet on Task 2.

variation across languages. The results are shown in Table 25.

The language-wise results reveal distinct differences between binary hate detection and target classification. Bengali performs particularly strongly on Task 1, with F1-scores above 96% for both classes. In contrast, its Task 2 performance is lower, particularly for the Religion category (57.14% F1), indicating increased difficulty in identifying the specific target of hateful content. English achieves comparatively strong target classification, particularly for Religion (93.62%) and Political (88.89%), while the Others category remains challenging across languages. Hindi shows a more balanced distribution across target categories, although Political and Others remain relatively difficult.

Overall, these results demonstrate that binary hate detection is comparatively stable across languages, whereas fine-grained target classification is more sensitive to language and target category. The class-wise analysis therefore provides additional evidence that the main challenge in Indic hateful meme understanding lies not only in identifying whether a meme is hateful, but also in determining the specific target of the hateful content.

L Dataset Duplication Analysis

To further evaluate dataset independence, we conducted an explicit duplication analysis between IndicHM and existing multilingual hateful meme benchmarks, particularly BHM. Since both datasets partially originate from overlapping social-media ecosystems, we examined potential redundancy at both the visual and textual levels.

Image-level overlap detection: We first identified visually similar memes using perceptual hashing (pHash) combined with FAISS-based nearest-neighbor retrieval. For each meme image I_i, we computed a perceptual hash representation:

$$\mathbf{h}_i = \phi(I_i), \tag{26}$$

where $\phi(\cdot)$ denotes the perceptual hashing operator. Each image hash was converted into a binary embedding vector and indexed using FAISS for efficient large-scale retrieval.

Given two images I_i and I_j, visual similarity was measured using the Hamming-distance approximation over the binary pHash vectors:

$$D_{\text{img}}(i,j) = \|\mathbf{h}_i - \mathbf{h}_j\|_0 , \tag{27}$$

where $\|\cdot\|_0$ denotes the number of differing hash bits. Two memes were considered near-duplicates if

$$D_{\text{img}}(i,j) < \tau_{\text{img}}, \tag{28}$$

where τ_{img} denotes a conservative image-similarity threshold.

Text-level overlap detection: To detect textual redundancy, we extracted OCR-normalized meme text and encoded each sample using SentenceTransformer embeddings:

$$\mathbf{z}_i = f_{\text{text}}(t_i), \tag{29}$$

where t_i denotes the OCR text associated with meme i, and $f_{\text{text}}(\cdot)$ represents the pretrained `all-mpnet-base-v2` encoder.

All textual embeddings were ℓ_2-normalized and indexed using FAISS inner-product search. Similarity between two textual samples was computed using cosine similarity:

$$S_{\text{text}}(i,j) = \frac{\mathbf{z}_i^\top \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \|\mathbf{z}_j\|_2}. \tag{30}$$

Two textual samples were treated as near-duplicates when

$$S_{\text{text}}(i, j) > \tau_{\text{text}}, \tag{31}$$

where τ_{text} denotes a strict similarity threshold.

Results and discussion: Using these complementary duplicate-detection strategies, we observed approximately 1.85% image-level overlap and 1.23% text-level overlap between IndicHM and BHM. These findings indicate that the datasets remain largely distinct despite partially overlapping source platforms.

The limited overlap is expected because BHM focuses exclusively on Bengali hateful memes, whereas IndicHM spans English, Hindi, and Bengali with substantially different collection keywords, communities, and annotation procedures. To ensure complete dataset independence, all detected overlapping samples will be removed before the final public release.