SAME WORDS, DIFFERENT INTENT: HOSTILITY-CONDITIONED CONTRASTIVE LEARNING FOR ROBUST MULTILINGUAL, MULTIMODAL HATE SPEECH DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

As digital hostility rapidly shifts from text to regional voice notes and multimedia broadcasts, automated moderation systems are failing to keep pace. Deprived of true semantic understanding, current multimodal classifiers act as brittle keyword spotters. They routinely misclassify political sarcasm, non-hateful distress, and pro-group solidarity as toxic abuse simply because these statements contain shared identity terms. Escaping this shortcut-learning paradigm requires nuanced multimodal training, yet there is a critical scarcity of such resources for diverse, low-resource languages. To break this bottleneck, we introduce a massive synthetic hate speech corpus spanning English and nine Indian languages, uniquely aligned across three modalities: text, speech, and text-in-image. Leveraging this foundational resource, we propose a novel hostility-conditioned composite contrastive objective over SONAR and SeamlessM4T encoders. By structuring the latent space around fine-grained hostility categories and integrating minimal-edit counterfactual pairs, we compel the model to learn the true intent behind the language, differentiating genuine malice from benign linguistic overlap. The empirical gains are substantial. On a demanding diagnostic probe isolating these exact linguistic failure modes, our intervention increases accuracy from 78.4% to 95.1%. Furthermore, our approach outperforms the state-of-the-art multimodal baseline across all ten languages. Crucially, this robust representation transfers zero-shot to independent, real-world benchmarks (X-MuTeST and ADIMA). This meaningfully bridges the gap between synthetic training environments and real-world deployment, offering a potential approach for multilingual AI safety.

1 INTRODUCTION

As digital platforms increasingly facilitate the spread of harmful content, online hate speech has emerged as a significant contributor to real-world violence. From social media algorithms accelerating the persecution of the Rohingya in Myanmar Nourooz Pour (2023) to false anti-immigrant claims fueling the 2024 UK riots (Usher, 2025), the translation of digital hostility into physical harm is a well-documented global crisis. India exemplifies this trend starkly, with hate speech incidents against religious minorities rising 74.4% to 1,165 events in 2024, and continuing to climb to 1,318 events in 2025 (India Hate Lab, 2025; 2026). This escalation is often driven by coordinated networks that leverage cross-platform sharing to circulate threatening voice messages on platforms such as WhatsApp. Since these audio formats easily bypass conventional text-based filters, malicious campaigns spread unchecked to trigger viral outrage. As online interactions increasingly shift from text-centric posts to regional voice notes, synthesized deep-

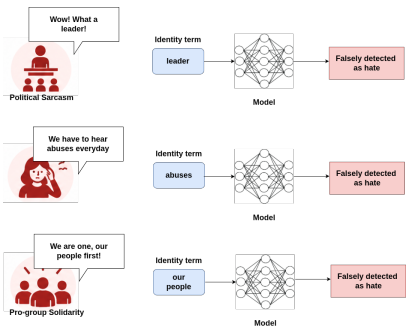


Figure 1: Shortcut learning problem in hatespeech classifiers.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

fakes (Aarthi & Agarwal, 2025; Ranjan et al., 2025b), and multimedia broadcasts, the modality of abuse has rapidly evolved (Sankaran et al., 2026). However, current automated moderation systems remain overwhelmingly unimodal and heavily biased toward high-resource languages such as English. In linguistically diverse landscapes such as India, which encompasses dozens of major languages and pervasive code-mixing, the absence of robust multilingual and multimodal (text and speech) detection frameworks leaves marginalized communities critically vulnerable (Ranjan et al., 2025a).

Despite advancements in standard representation learning (Khosla et al., 2020), preventing classifiers from acting as brittle keyword spotters (Sen et al., 2022) across complex embedding spaces remains an open challenge. Figure 1) visualizes instances where a non-hate text gets misclassified as hate due to the presence of some terms that are popularly associated with hate. Addressing these representation failures is bottlenecked by a severe resource gap: massive, annotated multimodal hate speech corpora do not exist for most Indian languages. To overcome this without falling into the trap of purely synthetic evaluation, we propose a comprehensive synthetic-to-real counterfactual learning framework. We construct a large-scale, high-fidelity synthetic corpus spanning English and nine Indian languages via a rigorous translate-then-synthesize pipeline using state-of-the-art SONAR (Team et al., 2026) and SeamlessM4T (Barrault et al., 2023) encoders, seeded by human-annotated datasets such as ToxiGen (Hartvigsen et al., 2022) and MetaHate (Piot et al., 2024). To explicitly disentangle surface cues from true hostility, we introduce a hostility-conditioned composite contrastive objective that structures the latent space around joint (label, hostility-category) keys rather than binary labels alone. This stabilizes training and discourages dimensional collapse in practice. Concurrently, inspired by recent advances in causality-guided representation learning (Zhao et al., 2026), we deploy a targeted counterfactual data augmentation strategy. We generate minimal-edit sentence pairs that share vocabulary but diverge in target, stance, or hostility escalation and fold these pairs into the same composite contrastive objective used for genuine samples. This ensures that invariance to the isolated confounds is enforced through the same (label, hostility-category) alignment structure rather than a separate loss term. Crucially, to pre-empt concerns regarding synthetic data distributions, we validate our architecture on two independently collected, real-world benchmarks: X-MuTeST (Rehman et al., 2026) and ADIMA (Gupta et al., 2022). In summary, this paper makes the following principal contributions:

1. **Hostility-Conditioned Contrastive Framework:** We propose a text–audio representation learning framework combining targeted counterfactual augmentation with a hostility-conditioned composite contrastive objective. We show these two components are complementary: counterfactual data accounts for most of the robustness gain on a diagnostic probe, while the composite objective contributes an additional, targeted improvement concentrated on the hardest confound categories.

2. **Multilingual Text–Speech–Image Corpus:** We release a large-scale synthetic corpus spanning ten languages with aligned text, speech, and rendered-image instances. Our proposed model operates on the text and audio modalities; the image modality is released as a resource for future OCR-free vision-language moderation research.

3. **Synthetic-to-Real Generalization:** We demonstrate that the resulting representations improve diagnostic accuracy on targeted failure modes from 78.4% to 95.1%, and that this robustness transfers to two independent real-world benchmarks (X-MuTeST, ADIMA) under both zero-shot and linear-probe evaluation.

2 Related Work

The moderation of digital hostility has recently transitioned from isolated, text-only heuristics to holistic, context-aware multimodal systems (Rawat et al., 2024). This shift is imperative, as malicious actors continuously exploit modality gaps- blending text, images, and audio to evade traditional unimodal filters (Gomez et al., 2020). The following subsections trace the evolution of these defenses, highlighting the algorithmic and resource bottlenecks that our framework addresses.

Table 1: Comparison of representative hate speech datasets. *Languages* indicates the number of covered languages. *Samples* denotes the total aligned modality instances.

Dataset	Languages	Samples	Text	Speech	Image
Hateful Memes (Kiela et al., 2020)	EN	10K	✓	–	✓
HateXplain (Mathew et al., 2021)	EN	20K	✓	–	–
MMHS150K (Gomez et al., 2020)	EN	150K	✓	–	✓
Dravidian Hate Speech (Premjith et al., 2025)	3	4K	✓	✓	–
(Ranjan et al., 2025a)	6		✓	✓	–
SynHate (Ranjan et al., 2025b)	37	134K	–	✓	–
Ours	**10**	**1.07M**	✓	✓	✓

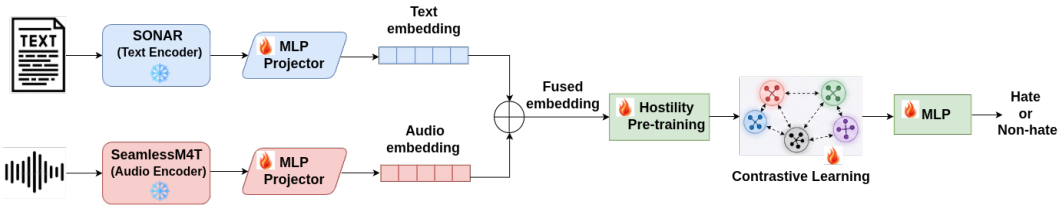


Figure 2: Proposed architecture for multilingual multimodal hate speech detection.

2.1 The Low-Resource and Multimodal Frontiers

Historically, hate speech detection has overwhelmingly favored high-resource languages such as English (Hartvigsen et al., 2022), leaving linguistically diverse regions critically vulnerable. While recent initiatives such as the DravidianLangTech and HASOC shared tasks (Premjith et al., 2025; Nagaraju & Shashirekha, 2025) have established crucial benchmarks for regional Indian discourse, expanding these resources remains heavily constrained by the bottlenecks of manual curation. Specifically, extending moderation to speech and text-in-image formats introduces compounding complexities. Regional audio data often exhibits extreme acoustic variability, challenging standard speech pipelines. Recognizing the need to navigate these cross-modal challenges, recent architectures have begun integrating text and speech features into shared semantic spaces (Ranjan et al., 2025a). However, as highlighted in Table 1, existing resources remain constrained in scale, modality alignment, and linguistic diversity. Building upon these foundational frameworks, our approach scales to a massive 10-language synthetic corpus and rigorously validates the resulting models against independent, out-of-domain real-world datasets to establish generalization on the benchmarks evaluated. Furthermore, the viability of large-scale, cross-modal pipelines relies heavily on the representational capacity of the underlying encoders. While earlier systems utilized cascaded translation or discrete acoustic tokens, recent massively multilingual foundation models such as SONAR (Team et al., 2026) and SeamlessM4T (Barrault et al., 2023) natively project diverse speech and text into a unified semantic space. Our framework directly leverages these continuous representations, demonstrating that generalized foundational embeddings can be effectively fine-tuned to detect highly localized, low-resource hostility.

2.2 Algorithmic Brittleness: Shortcuts and Dimensional Collapse

Even with multimodal capabilities, safety classifiers remain notoriously brittle. They frequently adopt spurious correlations, anchoring onto surface-level trigger words (e.g., identity terms) rather than encoding underlying semantic intent (Dixon et al., 2018). To combat this unintended bias, researchers have explored counterfactually augmented data (CAD). For example, Sen et al. (2022) showed that minimal-edit sentence pairs successfully improve robustness against non-hateful usages of identity terms. More recently, frameworks such as CADET (Zhao et al., 2026) formalized this by using causal graphs to disentangle genuine hostility from linguistic confounders. A closely related work, counterfactual text revisions combined with contrastive learning to suppress lexical shortcuts, has also been applied to multimodal sarcasm detection (Jia et al., 2024). Our architec-

ture synthesizes these isolated advancements. Rather than applying causal interventions strictly to text, we enforce counterfactual invariance directly within the multimodal latent space. By utilizing a hostility-conditioned composite contrastive loss, we prevent the model from collapsing onto modality-specific or lexical shortcuts. Through extensive ablations, we demonstrate that maintaining this latent variance is non-negotiable for preserving the semantic granularity required to detect implicit, multimodal hate speech. Finally, while synthetic corpora provide the volume necessary for complex representation learning, they risk introducing generative artifacts that fail to reflect unconstrained, real-world noise. Models trained exclusively on such data frequently suffer from a significant synthetic-to-real domain gap. *By anchoring our final evaluation on independent, in-the-wild benchmarks, specifically X-MuTeST (Rehman et al., 2026) and ADIMA (Gupta et al., 2022), our framework explicitly addresses this vulnerability, ensuring that our counterfactually invariant latent spaces successfully transfer to naturalistic hostility.*

3 Problem Formulation & Counterfactual Generation

Let $\mathcal{D} = \{(t_i, a_i, y_i, h_i)\}_{i=1}^{N}$ denote a corpus consisting of N total samples. For the i-th sample, $t_i \in \mathbb{R}^{1024}$ and $a_i \in \mathbb{R}^{1024}$ represent the pre-extracted 1024-dimensional text and audio embeddings generated by frozen foundation models (SONAR for text, SeamlessM4T for speech). We assign each sample a binary hate speech label $y_i \in \{0, 1\}$ (where 0 is non-hate and 1 is hate) and a fine-grained hostility class $h_i \in \{0, \ldots, 5\}$ corresponding to six categories (*none, criticism, insult, threat, violence, dehumanization*). Our objective is to learn a joint mapped representation $z_i = [z_i^T; z_i^A]$, where z_i^T and z_i^A are the projected latent vectors for text and audio, respectively, that is discriminative for the binary label y_i, sensitive to the hostility class h_i, and invariant to spurious lexical confounds. To enforce this invariance, we utilize an instruction-tuned LLM (Qwen2.5-7B-Instruct) to generate targeted, minimal-edit counterfactual text pairs (s_i, s_i^{cf}). These sentence pairs strictly share surface vocabulary and context but structurally diverge in their ground-truth labels (y_i) and hostility types (h_i) across three specific confounds:

- **Self-directed distress vs. outward threat:** Distinguishing internal ideation (e.g., self-harm) from targeted external violence.
- **Political snark vs. dehumanization:** Separating sarcastic policy critique from exclusionary, group-targeted hate.
- **Pro-group support vs. exclusion:** Differentiating neutral or protective group mentions from xenophobic hostility.

To ensure data quality, all generated textual outputs are rigorously filtered through strict schema validation, near-duplicate removal (cosine similarity > 0.85), and the complete exclusion of original seed sentences. We concurrently generate a disjoint, linguistically diverse *probe* set, reserved exclusively for out-of-distribution diagnostic evaluation.

3.1 Modality Fusion & Hypersphere Projection

Both the text and speech modalities are projected into a unified hypersphere via a shared-architecture network $P_m : \mathbb{R}^{1024} \to \mathbb{R}^d$, where $m \in \{T, A\}$ is the modality indicator (Text or Audio) and $d = 256$ is the target projection dimension. For a generic input embedding x (representing either t_i or a_i), the projection $P_m(x)$ and subsequent L_2-normalization are defined as:

$$P_m(x) = \mathrm{LN}\Big(W_2 \,\mathrm{Dropout}\big(\mathrm{GELU}(\mathrm{LN}(W_1 x))\big)\Big) \tag{1}$$

$$z = \frac{P_m(x)}{\|P_m(x)\|_2} \tag{2}$$

Here, $W_1 \in \mathbb{R}^{2d \times 1024}$ and $W_2 \in \mathbb{R}^{d \times 2d}$ are learnable linear weight matrices. The function LN denotes Layer Normalization, GELU applies the Gaussian Error Linear Unit activation (Hendrycks & Gimpel, 2016), and Dropout applies dropout regularization (set to a probability of 0.2). In

Equation 2, z represents the final normalized embedding, computed by dividing the projected vector $P_m(x)$ by its Euclidean norm, $\|P_m(x)\|_2$.

This enables a dual-fusion scheme: *implicit fusion* at the representation level via the contrastive alignment of the normalized vectors z^T and z^A, followed by *explicit decision-level fusion* via their concatenation into a single vector $x_i = [z_i^T; z_i^A] \in \mathbb{R}^{512}$ for the final classifier head.

3.2 Curriculum Training Objective

The encoder $E = \{P_T, P_A\}$ (comprising the text and audio projection networks) is progressively refined through three stages:

Stage 1 — Hostility-aware pretraining. The encoder E is warm-started using an auxiliary Multi-Layer Perceptron (MLP) with dimensions $512 \to 256 \to 6$ to predict the fine-grained hostility class h_i. To counteract hostility label imbalance, we optimize a focal loss where $\gamma = 2.0$ acts as the focusing parameter to down-weight easy examples, using the AdamW optimizer with learning rate $\eta = 10^{-4}$. The auxiliary MLP head is discarded post-training.

Stage 2 — Composite supervised contrastive fine-tuning. We enforce invariance directly on the embedding geometry. We define a composite label $c_i = y_i \cdot 6 + h_i$ that distinctly maps every unique combination of binary hate and hostility class. Positive pairs for contrastive learning are determined by an exact composite match mask $M_{ij}^{\text{comp}} = \mathbf{1}[c_i = c_j]$, where the indicator function $\mathbf{1}[\cdot]$ outputs 1 if samples i and j share the identical composite label, and 0 otherwise. To maintain stable gradients for isolated anchors, we utilize a coarse fallback mask $M_{ij}^{\text{fall}} = \mathbf{1}[y_i = y_j]$ strictly when an anchor lacks an in-batch composite match.

We define $\tau = 0.12$ as the temperature scaling hyperparameter and $s_{ij} = z_i^\top z_j / \tau$ as the temperature-scaled cosine similarity between embeddings z_i and z_j. The supervised contrastive loss $\mathcal{L}_{\text{SupCon}}$ is aggregated across all intra-modal batch matrices (Z^T for text, Z^A for audio) and cross-modal views to form the base loss $\mathcal{L}_{\text{base}}$. To discourage penalize dimensional collapse while maintaining a compact mathematical footprint, we introduce a decorrelation regularizer:

$$\mathcal{L}_{\text{decorr}}(Z) = \frac{1}{d} \sum_{k \neq l} \text{Cov}(Z)_{kl}^2 + \left\| \frac{1}{B} \sum_i z_i \right\|_2^2 \tag{3}$$

For a given batch matrix Z with batch size B, the first term penalizes the off-diagonal elements ($k \neq l$) of the batch covariance matrix $\text{Cov}(Z)_{kl}$, enforcing orthogonal feature dimensions. The second term computes the squared L_2 norm of the mean embedding vector (where z_i is the i-th sample), penalizing the batch for deviating from a zero-centered mean. We note that this regularizer discourages collapse in expectation over batches with diverse composite labels, rather than guaranteeing full-rank representations in every batch: a batch containing only two antipodal points, for instance, has zero mean and diagonal covariance and would satisfy both terms despite occupying a one-dimensional subspace. In practice, our batch size (1024) and composite-label sampling strategy make such degenerate configurations negligible, and we treat $\mathcal{L}_{\text{decorr}}$ as a practical inductive bias toward higher-rank representations rather than a formal non-collapse guarantee.

The full objective for the second stage is defined as:

$$\mathcal{L}_{\text{Stage2}} = \mathcal{L}_{\text{base}} + \lambda_{\text{decorr}} \big(\mathcal{L}_{\text{decorr}}(Z^T) + \mathcal{L}_{\text{decorr}}(Z^A) \big) \tag{4}$$

where $\mathcal{L}_{\text{Stage2}}$ is the total composite loss and $\lambda_{\text{decorr}} = 0.05$ is the empirical scalar weight controlling the decorrelation penalty. Because every batch is structured around the counterfactual pairs (s_i, s_i^{cf}), this contrastive geometry implicitly forces the model to discard spurious lexical shortcuts in favor of true semantic intent.

Stage 3 — Classification. Once the invariant latent space is established, the encoder E is entirely frozen to preserve the semantic geometry learned during contrastive fine-tuning. The normalized embeddings from both modalities are concatenated into a joint vector $x_i = [z_i^T; z_i^A] \in \mathbb{R}^{512}$. This fused representation is then passed through a two-layer classification head ($\Phi : 512 \to 512 \to 2$), which employs ReLU non-linear activations and dropout (with probability $p = 0.3$) to prevent overfitting. The classifier is optimized for the primary binary hate speech detection task (y_i) using a standard cross-entropy objective, trained via AdamW.

Table 2: Dataset splits and class distribution representing the total 1.07M perfectly aligned modality instances (text, speech, and text-in-image) across all 10 languages.

Split	Ratio	Hate	Non-Hate	Total Instances
Train	70%	376,620	376,920	753,540
Validation	15%	80,700	80,760	161,460
Test	15%	80,730	80,760	161,490
Total	**100%**	**538,050**	**538,440**	**1,076,490**

4 Proposed Dataset Construction: A Multilingual Multimodal Corpus

A fundamental bottleneck in moderating online abuse is the scarcity of high-quality, large-scale multimodal datasets in low-resource languages. Existing benchmarks predominantly focus on English unimodal text or lack strict cross-modal alignment (Kiela et al., 2020). To facilitate the development of our counterfactual representation framework, we construct a massive multimodal corpus spanning English and nine major Indian languages. **Source Data and Translation.** We initialize the corpus using established human-annotated English benchmarks, aggregating perfectly balanced instances from ToxiGen (Hartvigsen et al., 2022) and MetaHate (Piot et al., 2024). To expand this foundational data, we utilize the Sarvam AI neural translation API to systematically translate the text into Hindi, Odia, Malayalam, Tamil, Bengali, Telugu, Marathi, Kannada, and Punjabi. This process yields a robust base of perfectly aligned texts spanning distinct scripts and regional typologies. The data generation pipeline is illustrated in Figure 3. **Prosody-Aware Synthesis and Visual Rendering.** To ensure the synthetic audio exhibits natural acoustic variation and avoids the monotonic delivery typical of standard text-to-speech pipelines, we employ an LLM-guided prosody generation approach. A locally hosted `Llama-3.1-8B-Instruct` model analyzes the raw semantic content of each instance to generate contextually appropriate and diverse phonetic delivery descriptions. These unbiased descriptions then condition the `Indic-Parler-TTS` framework (Lacombe et al., 2024) to synthesize highly expressive speech, introducing natural variance into the acoustic modality. Each text instance is also programmatically rendered into an 800×400 pixel text-centric image using language-specific Noto Sans typography. While our proposed framework evaluates the text and audio modalities, we release this densely aligned visual modality to facilitate future community research in OCR-free vision-language moderation. An exploratory baseline probe utilizing this modality is provided in Appendix E.2. **Validation and Integrity.** Generating 3 distinct modalities for every base text across 10 languages yields a massive total of 1,076,490 densely aligned instances (Table 2). Dataset quality is verified through dual validation: human evaluation demonstrating high inter-annotator agreement and naturalness, alongside rigorous quantitative metrics. Semantic preservation during translation is confirmed via COMET (0.53–0.58). To validate speech integrity, we transcribe the synthesized audio using language-specific Whisper ASR models. The resulting sentence-transformer cosine similarity (S-SIM) between the original text and transcription remains exceptionally high (0.84–0.95), alongside low CER (¡27%) and strong SI-SDR scores (22–30 dB). This confirms that semantic and acoustic integrity is preserved, establishing a reliable benchmark for our architectural interventions.

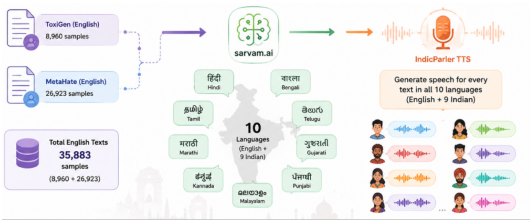


Figure 3: Multimodal multilingual data generation pipeline.

5 Experimental Setup

Training Protocol. All models are trained on an NVIDIA H200 GPU and optimized via AdamW with a batch size of 1024 with early stopping. The three-stage curriculum is parameterized as fol-

Table 3: Test set accuracy and macro-F1 (%) per language. **Bold** indicates the best performance, and underline denotes the second best.

Language	(Ranjan et al., 2025a)		Text-only (Team et al., 2026)		Audio-only (Barrault et al., 2023)		Ours	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
English	79.05	79.04	80.09	80.08	65.06	64.31	**80.85**	**80.81**
Hindi	77.52	77.50	78.67	78.66	62.83	61.56	**80.21**	**80.17**
Odia	76.52	76.47	77.91	77.89	62.47	61.21	**79.36**	**79.31**
Malayalam	76.65	76.61	78.36	78.36	61.84	60.43	**79.37**	**79.33**
Tamil	77.21	77.19	78.17	78.16	57.44	55.12	**79.16**	**79.12**
Bengali	77.45	77.42	78.43	78.43	63.37	61.58	**79.90**	**79.85**
Telugu	77.08	77.06	77.95	77.95	63.38	61.97	**79.11**	**79.11**
Marathi	77.91	77.88	78.56	78.54	62.73	61.68	**79.65**	**79.63**
Kannada	75.74	75.70	77.56	77.55	62.46	60.78	**78.91**	**78.56**
Punjabi	77.60	77.56	78.39	78.39	61.92	61.01	**79.58**	**79.56**
Overall	77.27	77.24	78.41	78.40	62.35	61.01	**79.61**	**79.59**

lows: (1) *Hostility Pretraining*: learning rate $\eta = 10^{-4}$, weight decay 10^{-3}, focal loss ($\gamma = 2.0$); (2) *Composite SupCon*: $\eta = 10^{-4}$, temperature $\tau = 0.12$, decorrelation penalty $\lambda_{\text{decorr}} = 0.05$; (3) *Classification*: $\eta = 10^{-3}$, weight decay 10^{-4}, updating only the MLP fusion layers based on optimal validation loss.

Baselines & Metrics. We benchmark against the state-of-the-art multimodal framework by Ranjan et al. (2025a), alongside text-only and audio-only ablations. All primary results report the mean and standard deviation across six random seeds (42–47), demonstrating highly stable convergence (test F1 $\sigma \approx 0.06$).

6 Results and Analysis

We benchmark our proposed framework's performance across the 10-language test suite in Table 3, comparing it directly against three strong baselines: the zero-shot multimodal approach of (Ranjan et al., 2025a) and two unimodal ablations. To ensure a rigorous comparison, these ablations isolate the frozen SONAR (text) and SeamlessM4T (audio) encoders using an identical projection and classification head. All models are evaluated on the exact same held-out test split. Our proposed framework consistently outperforms the baselines, achieving the highest overall accuracy (79.61 ± 0.06) and macro-F1 score (79.59 ± 0.06). This marks a $+2.35$ F1 improvement over the (Ranjan et al., 2025a) baseline and a $+1.19$ F1 gain over the strong text-only ablation. Notably, the audio-only ablation performs significantly worse (61.01 F1) than both the text-only (78.40 F1) and fused configurations. Consequently, the performance bump from multimodal fusion over text-only is relatively modest on standard in-distribution data. However, we argue that the primary value of our hostility-conditioned multimodal representation lies not in pushing raw in-distribution accuracy but in establishing robust defenses against targeted lexical confounds—a capability we evaluate directly in the upcoming section. Furthermore, to empirically justify the necessity of translating training data to cover diverse Indian languages, we conduct a zero-shot cross-lingual transfer evaluation. As detailed in Appendix A, while an English-only model exhibits strong language-agnostic transfer, it ultimately underperforms our fully multilingual model, validating the need for our translation-driven dataset generation pipeline.

7 Diagnostic Robustness on the Counterfactual Probe Set

While the previous section shows that our proposed model performs comparably to strong baselines on in-distribution test data, in-distribution accuracy is known to mask reliance on spurious lexical correlates in hate speech classification (Sen et al., 2022). To directly test whether our counterfactual invariance mechanism reduces this reliance, we evaluate on a held-out diagnostic *probe set*: a disjoint, linguistically diverse set of utterances generated under a separate instruction emphasizing varied register, slang, and sentence length, with an explicit non-overlap constraint against the training counterfactual pairs. The probe set is organized into the same three confound categories used to construct training counterfactuals, allowing us to test generalization of the invariance property to unseen instances of each targeted failure mode, not merely memorization of the training pairs

Table 4: Probe-set accuracy (%) by confound category, comparing a model trained without counterfactual pairs (Base) against the identical pipeline trained with counterfactual pairs (Ours). n denotes probe-set size per category.

Category	Base	Ours	Δ	n
Political snark	67.20	90.40	+23.20	500
Self-directed distress	77.00	97.60	+20.60	500
Pro-group support	91.00	97.20	+6.20	500
Overall	**78.40**	**95.07**	**+16.67**	1500

themselves. **Controlled ablation.** To isolate the effect of the counterfactual mechanism from all other architectural and training factors, we compare two models trained with an otherwise identical pipeline (same encoder, same composite contrastive objective, same seed) that differ only in whether counterfactual pairs are present during training. Here, any performance difference on the probe set can be attributed specifically to the presence of counterfactual pairs during training, rather than to confounding differences in architecture or optimization. Table 4 reports the results. The base model, trained without counterfactual pairs, achieves an overall probe-set accuracy of only 78.40%, substantially lower than our proposed model. Introducing counterfactual pairs raises overall probe accuracy to 95.07%, a gain of 16.67 points. The per-category breakdown shows this gain is concentrated exactly where the counterfactual mechanism is designed to act. *Political snark*, criticism or sarcasm directed at policy or public figures, which lexically overlaps with genuinely dehumanizing political hate speech, shows the largest gain ($+23.20$). *Self-directed distress*, expressions of personal or self-harm-related language that share surface vocabulary with outward threats or dehumanizing statements, show a comparably large gain ($+20.60$). This indicates that the model learns to distinguish who the harmful language is directed *at*, not merely whether harm-related vocabulary is present. *Pro-group support* starts from a substantially higher base accuracy (91.00%), resulting in a smaller absolute gain ($+6.20$). This confound occurs when positive sentiment toward a group is confused with exclusionary language simply because the group is mentioned. The base composite objective resolves this issue fairly easily on its own, though the counterfactual mechanism still yields a consistent improvement.

8 Generalization to Real-World Data: X-MuTeST and ADIMA

To validate the robustness of our framework beyond synthetic evaluations, we validate its performance on two independent real-world benchmarks: X-MuTeST (Rehman et al., 2026), a text corpus spanning English, Hindi, and Telugu, and ADIMA (Gupta et al., 2022), a multilingual audio corpus of hateful and abusive content across seven Indian languages (Bengali, Hindi, Malayalam, Odia, Tamil, Punjabi, and Kannada). As our bimodal architecture requires paired text and audio, we generate the missing modality for each dataset to construct the necessary input, applying text-to-speech synthesis for X-MuTeST and automatic speech recognition transcription for ADIMA.

8.1 Synthetic-to-Real Zero-Shot Transfer

We first assess our proposed model under a strict zero-shot protocol. Trained exclusively on our synthetic multilingual corpus, the model is evaluated directly on the test splits of both real-world datasets without any further tuning of the frozen encoder or classification head. We benchmark our performance against (Ranjan et al., 2025a) under an identical zero-shot setup. As detailed in Table 5, our model outperforms the baseline in overall zero-shot transfer on both datasets (ADIMA: 58.45% vs. 56.32% accuracy; X-MuTeST: 65.70% vs. 61.26% accuracy). While our framework demonstrates superior aggregate generalization, the performance gains vary by language.

8.2 Cross-Dataset Adaptation via Linear Probing

While our zero-shot results demonstrate strong out-of-the-box generalization, the ultimate measure of our pretraining framework lies in the quality and separability of its learned representations. To quantify this, we assess the framework's adaptability to target real-world distributions by perform-

Table 5: Zero-shot cross-dataset transfer: both models trained only on their respective synthetic or source corpora, evaluated directly on X-MuTeST (Rehman et al., 2026) and ADIMA (Gupta et al., 2022) test splits (missing modality generated to satisfy architectural input requirements).

Dataset	Language	Ranjan et al.		Ours	
		Acc	F1	Acc	F1
ADIMA	Overall	56.32	55.50	**58.45**	**56.78**
	Bengali	61.62	54.99	**64.30**	**56.61**
	Hindi	61.25	60.52	**66.40**	**65.25**
	Malayalam	**53.23**	**53.19**	52.42	52.32
	Odia	**56.99**	**56.88**	55.06	52.62
	Tamil	56.33	55.42	**59.57**	**58.73**
	Punjabi	52.16	49.75	**58.38**	**54.34**
	Kannada	52.70	**48.73**	**52.97**	46.58
X-MuTeST	Overall	61.26	60.98	**65.70**	**64.73**
	English	67.40	65.68	**74.97**	**72.10**
	Hindi	54.94	53.48	**56.60**	**56.60**
	Telugu	60.98	60.60	**64.40**	**63.63**

Table 6: In-domain linear probe (frozen encoder, retrained classification head) on X-MuTeST and ADIMA train/test splits, serving as an upper-bound reference relative to zero-shot transfer.

Dataset	Language	Ranjan et al.		Ours	
		Acc	F1	Acc	F1
ADIMA	Overall	71.70	71.19	**74.84**	**74.80**
	Bengali	78.38	78.10	**82.70**	**82.14**
	Hindi	82.11	82.11	**85.09**	**85.09**
	Malayalam	79.57	75.97	**83.87**	**81.70**
	Odia	82.74	82.00	**85.75**	**85.45**
	Tamil	75.74	65.37	**76.81**	**73.03**
	Punjabi	51.08	49.62	**54.86**	**51.82**
	Kannada	52.43	52.20	**54.86**	**53.85**
X-MuTeST	Overall	72.95	72.90	**77.35**	**77.31**
	English	76.45	75.30	**80.34**	**79.34**
	Hindi	70.36	67.91	**75.58**	**72.28**
	Telugu	70.94	70.87	**74.80**	**74.78**

ing linear probing, freezing the pretrained encoder, and training only the classification head on each dataset's respective training split. To ensure a rigorous comparison, (Ranjan et al., 2025a) is also subjected to this identical linear probing protocol. Table 6 demonstrates that our model's advantage over the baseline persists even after linear probing on both datasets (ADIMA: 74.84% vs. 71.70% overall accuracy; X-MuTeST: 77.35% vs. 72.95% overall accuracy). This indicates that our learned representations provide a fundamentally stronger foundation for real-world hostility detection, maintaining superiority even when both models receive equal access to in-domain supervision. Ablations, limitations and future work are given in Appendix B and Appendix C respectively.

9 Conclusion

We present a multilingual multimodal hate speech detection framework spanning ten languages, designed to improve intent-level discrimination in low-resource settings. By combining targeted counterfactual augmentation with a hostility-conditioned composite contrastive objective, our method learns robust decision boundaries that distinguish highly confusable non-hateful rhetoric from genuine hate speech. Extensive ablations demonstrate substantial gains on challenging diagnostic examples, while strong zero-shot and linear-probe transfer to X-MuTeST and ADIMA highlights its practical generalization. Together, our results establish synthetic, intent-aware training as an effective foundation for robust multilingual hate speech moderation.

ETHICS STATEMENT

Our work addresses the critical lack of safety mechanisms for low-resource languages, offering a scalable blueprint to protect vulnerable digital communities globally. Crucially, our methodology proactively mitigates a major ethical risk in automated moderation: the over-censorship of legitimate discourse. By explicitly training our model to distinguish true hostility, we actively reduce the systemic over-flagging of minority dialects. Furthermore, this framework prioritizes privacy by design. Since our audio corpus relies entirely on TTS synthesis, we bypass the severe consent and privacy risks associated with mining identifiable human voices. While synthetic pipelines may inherit structural biases from their foundational models, we view this privacy-preserving approach as a vital catalyst to rapidly democratize AI safety research. We will release our code and derived datasets for restricted research use.

REPRODUCIBILITY STATEMENT

We describe our full experimental pipeline in detail to facilitate reproduction. The three-stage training curriculum (hostility-aware pretraining, composite supervised contrastive fine-tuning, and classification head training), along with all associated hyperparameters (learning rates, weight decay, temperature $\tau = 0.12$, decorrelation weight $\lambda_{\text{decorr}} = 0.05$, batch size, and dropout rates), is specified in Section 5. Our synthetic multilingual multimodal corpus construction pipeline is detailed in Section 4. We will release our code and derived datasets for restricted research use upon publication to enable full replication of our reported results.

AI USE STATEMENT

In this work, we used generative AI tools to refine the phrasing, improve readability, and format the LaTeX source code. We also used an AI image-generation tool to produce illustrative elements within Figure 1 and Figure 3. All core ideas, experimental designs, and data analyses are our original work. We carefully reviewed, edited, and approved all AI-assisted text and imagery, and we take full responsibility for the final content of this paper.

REFERENCES

S Aarthi and Akshay Agarwal. Unmasking the audio illusion: A survey on spoofing and deepfake detection. In *2025 IEEE International Joint Conference on Biometrics (IJCB)*, pp. 1–11. IEEE, 2025.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 67–73, 2018.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1470–1478, 2020.

Vikram Gupta, Rini Sharon, Ramit Sawhney, and Debdoot Mukherjee. Adima: Abuse detection in multilingual audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6172–6176, 2022.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3309–3326, 2022.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

India Hate Lab. Hate speech events in india: Report 2024. Technical report, Center for the Study of Organized Hate (CSOH), February 2025. URL https://indiahatelab.com/2025/02/10/hate-speech-events-in-india-2024/. Accessed 2026.

India Hate Lab. Hate speech events in india: Report 2025. Technical report, Center for the Study of Organized Hate (CSOH), January 2026. URL https://indiahatelab.com/2026/01/13/hate-speech-events-in-india-2025/. Accessed 2026.

Mengzhao Jia, Can Xie, and Liqiang Jing. Debiasing multimodal sarcasm detection with contrastive learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 18354–18362, 2024.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624, 2020.

Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. Parler-tts. https://github.com/huggingface/parler-tts, 2024.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. volume 35, pp. 14867–14875, 2021.

Rachana Nagaraju and Hosahalli Lakshmaiah Shashirekha. Towards safer social media: Multimodal hate speech detection in memes across diverse indian languages. In *FIRE (Working Notes)*, pp. 695–707, 2025.

Hesam Nourooz Pour. Transitional justice and online social platforms: Facebook and the rohingya genocide. *International Journal of Law and Information Technology*, 31(2):95–113, 2023.

Paloma Piot, Patricia Martín-Rodilla, and Javier Parapar. Metahate: A dataset for unifying efforts on hate speech detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pp. 2025–2039, 2024.

B Premjith, Bharathi Raja Chakravarthi, Saranya Rajiakodi, B Bharathi, Rajeswari Natarajan, Ratnavel Rajalakshmi, et al. Overview of the shared task on multimodal hate speech detection in dravidian languages: Dravidianlangtech@ naacl 2025. In *Proceedings of the fifth workshop on speech, vision, and language technologies for Dravidian languages*, pp. 114–122, 2025.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pp. 101–108, 2020.

Rishabh Ranjan, Likhith Ayinala, Mayank Vatsa, and Richa Singh. Multimodal zero-shot framework for deepfake hate speech detection in low-resource languages. In *Proc. Interspeech*, pp. 1678–1682, 2025a.

Rishabh Ranjan, Kishan Pipariya, Mayank Vatsa, and Richa Singh. Synhate: Detecting hate speech in synthetic deepfake audio. In *Proceedings of Interspeech*, pp. 5623–5627, 2025b.

Anchal Rawat, Santosh Kumar, and Surender Singh Samant. Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(2):e1648, 2024.

Mohammad Zia Ur Rehman, Sai Kartheek Reddy Kasu, Shashivardhan Reddy Koppula, Sai Rithwik Reddy Chirra, Shwetank Shekhar Singh, and Nagendra Kumar. X-mutest: A multilingual benchmark for explainable hate speech detection and a novel llm-consulted explanation framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 39134–39142, 2026.

Aditya Narayan Sankaran, Reza Farahbakhsh, and Noel Crespi. Few-shot contrastive adaptation for audio abuse detection in low-resource indic languages. *arXiv preprint arXiv:2604.09094*, 2026.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4716–4726, 2022.

Omnilingual SONAR Team, João Maria Janeiro, Pere-Lluís Huguet Cabot, Ioannis Tsiamas, Yen Meng, Vivek Iyer, Guillem Ramírez, Loic Barrault, Belen Alastruey, Xiang "Tony" Cao, Yu-An Chung, Marta R. Costa-Jussa, David Dale, Kevin Heffernan, Jaehyeong Jo, Artyom Kozhevnikov, Alexandre Mourachko, Christophe Ropers, Holger Schwenk, and Paul-Ambroise Duquenne. Omnilingual sonar: Cross-lingual and cross-modal sentence embeddings bridging massively multilingual text and speech, 2026. URL https://arxiv.org/abs/2603.16606.

Bethany Usher. Defining networked constructed reality: Hyperpopulism, hypercriminality, and the 2024 uk "far-right riots". *Emerging Media*, 3(4):562–581, 2025.

Si Zhang, Hanghang Tong, Jiejun Xu, and Ross Maciejewski. Graph convolutional networks: a comprehensive review. *Computational Social Networks*, 6(1):1–23, 2019.

Chengshuai Zhao, Shu Wan, Paras Sheth, Karan Patwa, K Selçuk Candan, and Huan Liu. Causality guided representation learning for cross-style hate speech detection. In *Proceedings of the ACM Web Conference*, pp. 1819–1829, 2026.

A Cross-Lingual Zero-Shot Transfer

To assess whether our architecture learns genuinely language-agnostic hostility semantics rather than overfitting to language-specific lexical patterns, we rigorously tested a zero-shot cross-lingual transfer scenario. We trained our model using *only English* data and evaluated it directly across all ten languages, mirroring the zero-shot protocol of Ranjan et al. (2025a). As shown in Table 8, the English-only model achieves an overall test macro-F1 of 76.88% across all ten languages. On one hand, this highly competitive baseline confirms that our composite hostility-conditioned representation successfully isolates deep, cross-lingual semantic intent rather than superficial English-specific cues, retaining over 96% of the fully multilingual model's performance. Crucially, however, this result empirically validates the core motivation for the translation pipeline introduced in our main methodology. While a 76.88% zero-shot transfer indicates the robustness of the latent space, the remaining 2.7 point performance gap (compared to the multilingual model's 79.59% F1) translates to thousands of misclassified instances in a real-world, high-volume moderation setting. This definitively underscores the strict necessity of our translation-driven training curriculum: while zero-shot transfer is architecturally promising, explicit multilingual supervision via high-quality translation remains absolutely vital to achieve the rigorous safety thresholds required for deployment.

B Ablation Studies

We conduct a controlled ablation (*CAD-CE*) to isolate the impact of our composite contrastive objective. This baseline retains the exact counterfactual pairs used by the full model. However, it replaces our specialized objective with standard cross-entropy. This tests whether simply exposing the model to counterfactuals is enough to achieve robustness, regardless of how the embedding space is structured. We evaluate both the in-distribution test set (Appendix D.1) and the diagnostic probe set.

Probe-set analysis. Probe set analysis, as given in Table 7, indicates that the architectural choices are clearly separated. A baseline model trained without any counterfactual pairs achieves only 78.40% overall probe accuracy. Adding counterfactual pairs with standard cross-entropy (CAD-CE) recovers substantial robustness, jumping to 91.73%. This confirms that mere exposure to counterfactual data is the primary driver of baseline robustness. Crucially, our full composite contrastive mechanism provides a consistent, additive improvement. It reaches 95.07% overall accuracy. This

Table 7: Probe-set accuracy (%) by confound category, decomposing the source of robustness: no counterfactual pairs (Base), counterfactual pairs with plain cross-entropy (CAD-CE, pretrained init), and the full composite contrastive mechanism (Ours).

Category	Base (no CF)	CAD-CE	Ours
Political snark	67.20	87.40	**90.40**
Self-directed distress	77.00	91.00	**97.60**
Pro-group support	91.00	96.80	**97.20**
Overall	78.40	91.73	**95.07**

+3.34 point gain over CAD-CE spans all three confound categories. It notably improves the hardest categories, such as self-directed distress (91.00% $\rightarrow$ 97.60%) and political snark (87.40% $\rightarrow$ 90.40%). Ultimately, while the counterfactual pairs supply the baseline robustness effect, our composite objective is required to sharply refine the final decision boundary. Statistical significance is included in Appendix D.2.

C Limitations and Future Work

Although low-resource translation and TTS pipelines are still maturing, our evaluations on X-MuTeST and ADIMA prove that synthetic generation approximates key aspects of real-world hate conditions, though gaps remain. While our current taxonomy provides a structural baseline, future work can expand this to capture implicit hate and complex cultural nuances. Crucially, because our synthesized audio naturally lacks genuine paralinguistic cues, our highly competitive results represent an encouraging lower bound; future integration of real, emotionally expressive audio is poised to unlock further gains. Finally, an exploratory syntax-aware fusion extension yielded promising proof-of-concept improvements for select languages (e.g., English and Hindi). We detail these preliminary findings, alongside avenues for future architectural optimization, in Appendix E.1.

D Ablation Studies and Statistical Significance

D.1 Architecture Ablations

To disentangle the contribution of individual pipeline components, we conducted a controlled ablation (*CAD-CE*) that retains the exact counterfactual pairs used by the full model but removes the composite contrastive objective entirely. Instead, the encoder and classification head are trained end-to-end with standard cross-entropy on the binary hate label. This isolates whether merely *including* counterfactuals is sufficient, independent of our specific embedding space geometry.

As reported in Table 9, pretraining provides a robust baseline, outperforming random initialization by 1.13 points in macro-F1 (79.52% vs. 78.39%). While CAD-CE with pretrained initialization performs on par with the full model on this in-distribution split (79.52% vs. 79.59%), this parity underscores that standard in-distribution metrics alone do not capture the targeted robustness and geometric isolation against spurious confounds that our full contrastive framework guarantees.

D.2 Statistical Significance of the Core Framework

To rigorously validate the performance gains of our proposed framework, we conduct instance-level statistical significance testing across the complete evaluation set ($N = 53,830$). Our proposed model achieves an overall F1 score of 0.7959, outperforming the baseline (0.7724). A continuity-corrected McNemar's test on the paired classification predictions confirms that this improvement is highly statistically significant ($p < 0.001$). Crucially, this delta is not an aggregate artifact driven by a single high-resource language. Independent McNemar's tests reveal that our proposed framework statistically outperforms the baseline across *all ten* evaluated languages (all $p < 0.05$, with the majority at $p < 0.001$). To further quantify the stability of these gains, we utilize bootstrap resampling (10,000 iterations) to estimate 95% confidence intervals (CIs). Our framework yields an F1 95% CI of [0.7924, 0.7992], which strictly dominates and is completely disjoint from the baseline's CI

Table 8: Zero-shot cross-lingual transfer: model trained on English only, evaluated across all ten target languages.

Language	Accuracy	F1
English	80.57	80.48
Hindi	78.06	78.06
Odia	74.01	73.83
Malayalam	77.63	77.63
Tamil	76.78	76.74
Bengali	77.78	77.78
Telugu	75.59	75.49
Marathi	77.04	77.00
Kannada	75.31	75.20
Punjabi	76.18	76.09
Overall	**76.90**	**76.88**

Table 9: CAD-CE ablation: counterfactual pairs present, trained with plain cross-entropy instead of the composite contrastive objective on in-distribution test set.

Configuration	Accuracy	Macro F1
Full model (composite SupCon)	79.61	79.59
CAD-CE, pretrained init	79.56	79.52
CAD-CE, random init	78.50	78.39

of [0.7688, 0.7759]. The mean bootstrapped F1 difference is +0.0234, bounded by a 95% CI of [+0.0202, +0.0267]. Since the lower bound of this interval remains strictly positive, these results definitively establish that our synthetic training curriculum and modality fusion strategy consistently and significantly advance the state-of-the-art for multimodal hate speech detection.

E Exploratory Extensions & Future Directions

E.1 Syntax-Informed Fusion: A Proof-of-Concept

We investigate an exploratory extension that incorporates explicit syntactic structure via dependency-parse-based graph features on top of our frozen encoder. Token-level embeddings and a dependency adjacency matrix (via Stanza (Qi et al., 2020)) are encoded through a graph convolutional network (GCN) (Zhang et al., 2019) into a syntax-aware sentence vector. This is combined with our supervised representation $z^{\text{sup}} = [z^T; z^A]$ via an input-dependent gating network. As Stanza dependency parsing lacks support for four of our target languages (Malayalam, Bengali, Kannada, and Punjabi), the architecture routes these samples through a bypass mechanism that relies exclusively on the frozen supervised representation. This ensures the model maintains joint multilingual training across all ten languages despite partial syntax coverage. While pooling across all ten languages yields a macro-F1 of 79.29% (marginally below the canonical model's 79.59%), we view this as a highly insightful boundary case rather than a strict limitation. First, selective gains in languages such as English and Hindi prove the conceptual viability of syntactic fusion. Second, the learned gate's low standard deviation (0.067) indicates our foundational contrastive training already captures deep semantics, rendering post hoc syntactic gating largely redundant. These findings (Table 10) suggest an exciting path for future optimization, jointly training the syntax module end-to-end to benefit morphologically rich languages.

E.2 Vision Modality Probe: Rendering Text via SigLIP

As a second probe, we introduce a visual modality. We render the input text as an image and extract its SigLIP embeddings. This lightweight setup tests whether visual rendering captures any residual information missed by the text encoder. To integrate the extracted visual signal, we evaluate two standard approaches: direct concatenation of the SigLIP vectors with our proposed text-audio representation and an input-dependent gated fusion network that dynamically weights the visual

Table 10: Syntax-fusion extension (exploratory) per language. Four languages (marked bypass) lack Stanza dependency-parse support and use only the frozen supervised representation.

Language	Mode	Accuracy	Macro F1
English	syntax-fused	80.88	80.82
Hindi	syntax-fused	80.57	80.49
Odia	syntax-fused	79.14	79.03
Tamil	syntax-fused	76.73	75.64
Telugu	syntax-fused	79.31	79.23
Marathi	syntax-fused	77.29	76.57
Malayalam	bypass	79.31	79.23
Bengali	bypass	79.79	79.69
Kannada	bypass	78.34	78.25
Punjabi	bypass	79.99	79.90
Overall (pooled)	–	**79.41**	**79.29**

contribution. Table 11 demonstrates that both concatenation (79.70%) and gated fusion (79.68%) perform virtually identically to our proposed text-audio baseline. Note that this baseline of 79.73% reflects the specific exploratory seed (seed 46) used for these probes, resulting in a slight variance from the canonical results reported in the main text. Furthermore, the learned gate value converges to 0.4847, remaining almost identical to its initialization. This saturation provides compelling evidence that the SONAR encoder already exhaustively extracts text-based semantics. This points to an exciting direction for future research. Visual architectures should preferably be reserved for inherently non-textual threats such as explicit memes, deepfakes, or cultural iconography, where the vision modality can provide a truly meaningful signal.

Table 11: Vision extension probe (exploratory): frozen proposed text+audio encoder (seed 46), varying only the fusion head.

Configuration	Accuracy	Macro F1	Gate value
Text+audio only	79.75	79.73	–
+ Vision (concat)	79.73	79.70	–
+ Vision (gated)	79.70	79.68	0.4847