ORBIT-CL: A Semantically-Aware and Lightweight Multimodal Transformer for Cyberbullying Detection

Saed Alqaraleh
Department of Data Science and Artificial Intelligence, College of Information Technology, Mutah University, Karak, Jordan
saed.alqaraleh@mutah.edu.jo (corresponding author)

Received: 14 February 2026 | Revised: 28 March 2026 | Accepted: 4 April 2026

 DOI: https://doi.org/10.48084/etasr.18187

ABSTRACT

Cyberbullying appears in complex forms, namely (1) syntactic obfuscations, (2) context-dependent multimodal memes, and (3) semantically ambiguous text (e.g., sarcasm or reclaimed slurs). The present study proposes a Contrastive and Lightweight Multimodal Transformer, ORBIT-CL, an end-to-end system addressing all three challenges. It combines lightweight multimodal feature extraction (Transformer-based OCR (TrOCR) and visual object tags) with a RoBERTa-based classifier. This classifier employs a dual-robustness objective: adversarial noise training alongside semantic contrastive learning to handle both syntactic attacks and intent ambiguity. The system uses multi-task heads for binary bullying detection and ordinal hostility and calibration modules to maintain stable moderation thresholds. The experimental protocol, which includes challenge sets, was designed to address semantic ambiguity. Using publicly available datasets (HateXplain and Hateful Memes), ORBIT-CL achieved highly competitive performance (0.88 Macro-F1) on the full HateXplain benchmark. To specifically validate its semantic robustness, it also achieved a 30-point F1 gain on a targeted challenge set of ambiguous content (e.g., sarcasm, reclaimed slurs), thereby addressing a key failure mode of prior models. Furthermore, by fusing lightweight visual tags as text, the introduced model achieved 0.895 Area Under the Receiver Operating Characteristic (AUROC) curve on Hateful Memes, demonstrating a highly efficient alternative to heavy multimodal baselines while retaining the efficiency of a text-only encoder. By unifying lightweight multimodal fusion with a dual (syntactic and semantic) robustness framework, ORBIT-CL contributes to deployable, context-aware, and efficient cyberbullying detection.

Keywords-cyberbullying; content moderation; OCR; RoBERTa; multimodal memes

I. INTRODUCTION

Cyberbullying is widespread among adolescents and young adults, correlating with worsened mental health outcomes and indicators of self-harm risk [1, 2]. With the evolution of communication platforms, especially the social and messaging ones, abusive content includes items beyond text, such as images and memes. While transformer models, such as RoBERTa and HateBERT, achieve high text-processing performance [3-6], they face significant operational gaps. They are highly vulnerable to syntactic perturbations and frequently misclassify semantically ambiguous text, such as sarcasm or reclaimed slurs. Furthermore, these text-only models inherently fail to process visual context. Despite the aforementioned gaps, existing multimodal solutions heavily rely on massive Vision-Language Models (VLMs), lacking deployable, lightweight alternatives. To close this performance gap between fragile text models and heavy multimodal systems, the present study proposes ORBIT-CL (a Contrastive and Lightweight Multimodal Transformer). ORBIT-CL builds on a robust Optical Character Recognition (OCR)-based pipeline through two key innovations: (1) a lightweight multimodal fusion strategy (C1) that enriches the text encoder with visual object tags, and (2) a dual-robustness objective (C2) based on syntactic adversarial training [7, 8] and a novel semantic contrastive loss to resolve intent ambiguity. Therefore, the novelty of this work lies not in proposing a new transformer architecture, but in a highly efficient engineering integration.

The main contributions of this work are: (1) a lightweight, multimodal fusion technique that incorporates visual object tags alongside text to efficiently process context-dependent hate speech; (2) a dual-robustness framework that resolves intent ambiguity by combining syntactic noise training with a semantic contrastive loss; (3) a unified OCR + Natural Language Processing (NLP) pipeline providing a reproducible protocol [9, 10]; and (4) in-depth experimental work featuring semantic challenge sets, ablations, and efficiency reporting. Transformer backbones for cyberbullying and hate/abuse detection systems have been investigated [11]. RoBERTa variants with feature fusion (e.g., GloVe) report gains on benchmark corpora [5]. Domain-specific pretraining, such as

HateBERT, improves abusive language detection [4]. Hybrid deep learning models, such as ProTect, combine Convolutional Neural Networks (CNNs) and Random Forests (RFs) to capture context and user cues [3]. Similarly, deep hybrid models have been used to ensure safer digital communication on social networks [12]. Authors in [13] showed that machine learning classification, especially ensemble-based methods, can achieve strong results when features are well engineered. Authors in [14] investigated the possibility of using Large Language Models (LLMs) for both synthetic data augmentation and direct classification.

In the multimodal domain, memes and screenshots containing text should be considered in high-risk contexts, such as hateful content [15]. However, efficient systems must consider both text and memes rather than relying on memes alone, since the latter can appear easy to classify when interpreted in isolation under the assumption of a single meaning; nevertheless, combining them with text may change the overall meaning [16]. While OCR and scene-text understanding have matured considerably [10] and recent Transformer-based OCR (TrOCR) achieves strong recognition performance without handcrafted decoders [9], effectively integrating both is significant. ORBIT leverages OCR to expand coverage to these cases while keeping the downstream detector text-centric.

II. METHODS

A. Problem Formulation

Given an input item x that is either a text string or an image containing text, ORBIT outputs: (i) a bullying label $y \in \{0,1\}$, (ii) a hostility level $h \in \{1,2,3\}$ (low/medium/high), and (iii) a calibrated score $s \in [0,1]$ for thresholding.

B. System Overview

The main steps of the proposed approach are (Figure 1):

1) Step 1: Input Processing

Regarding text, raw strings were passed, while for images, two parallel processes were considered:

1. Text extraction: OCR (TrOCR-base [9]) was run to extract candidate text boxes and transcriptions; low-confidence boxes were filtered out.
2. Visual feature extraction: A lightweight, pre-trained object detector (e.g., You Only Look at One Sequence (YOLOS)) extracted high-confidence visual keywords (e.g., ['person', 'dog', 'gun']) from the image. These tags provide essential semantic context; for example, a 'person' tag can distinguish a dehumanizing attack from a harmless comment about an object without the high computational cost of a full visual encoder.

2) Step 2: Text Normalization and Fusion

A multi-stream fusion strategy was employed. The encoder ingested three separate text streams, concatenated with special tokens:

1. The raw text (from Step 1-i, i.e., the original text input).
2. A normalized version of the raw text, using the lightweight normalizer (regexes + mappings) to handle obfuscations (e.g., 1 to i, and 0 to o).
3. The string of visual object tags from Step 1-2 (e.g., tags: person, dog, gun).

This lightweight, text-centric fusion allows the model to understand visual context without a heavy image encoder.

3) Step 3: Encoder and Heads

A RoBERTa-base encoder processed the fused three-part text input. It produced (i) a bullying head (binary) and (ii) an ordinal hostility head (cumulative link) [17]. Let $y \in \{0,1\}$ be the binary bullying label and $h \in \{1, \dots, K\}$ be the ordinal hostility level, where K=3 corresponds to low, medium, and high hostility. The dual-robustness objective was a weighted sum of four losses:

$$\mathcal{L}_{total} = \lambda_y \mathcal{L}_{focal} + \lambda_h \mathcal{L}_{ord} + \lambda_{cons} \mathcal{L}_{cons} + \lambda_{cl} \mathcal{L}_{cl} \quad (1)$$

where the λ terms are scalar weights. In all experiments, these were empirically set to $\lambda_y = 1.0$, $\lambda_h = 1.0$, $\lambda_{cons} = 0.1$, and $\lambda_{cl} = 0.1$, balancing the primary classification tasks with lightweight syntactic and semantic regularization.

a) Focal Loss (Binary)

For the binary task y, focal loss was used to downweight well-classified examples (easy samples). Let $\hat{p} \in [0,1]$ be the model's predicted probability for the positive class ($y = 1$):

$$\mathcal{L}_{focal}(y, \hat{p}) = -\alpha y (1 - \hat{p})^{\gamma} \log \hat{p} - (1 - \alpha)(1 - y) \hat{p}^{\gamma} \log(1 - \hat{p}) \quad (2)$$

b) Ordinal Head (Cumulative Link)

For the ordinal task h, a cumulative link model was used [17]. This model predicts $K - 1$ logits $\hat{q} = \{\eta_1, \dots, \eta_{K-1}\}$, representing the log-odds of $h > k$. A set of binary indicators $U_k = \mathbb{I}(h > k)$ was defined, where $\mathbb{I}(\cdot)$ is the Iverson bracket (1 if true, 0 otherwise). The loss is the sum of binary cross-entropy losses for each of the $K - 1$ links, where $\sigma(\cdot)$ is the logistic sigmoid function:

$$\mathcal{L}_{ord(h,\hat{q})} = -\sum_{k=1}^{K-1} \left(u_k \log \sigma(\eta_k) + (1 - u_k) \log \left(1 - \sigma(\eta_k)\right) \right) \quad (3)$$

c) Consistency Regularization (Syntactic Robustness)

Let $P_\theta(\cdot \mid x)$ be the model's output distribution for input x, and let $\tilde{x}$ be an augmented version of the input created via character-level noise. The loss is the Kullback–Leibler (KL) divergence between their outputs:

$$\mathcal{L}_{cons} = KL(p_\theta(\cdot \mid x) \| p_\theta(\cdot \mid \tilde{x})) \quad (4)$$

d) Semantic Contrastive Regularization

To solve semantic ambiguity failures identified in preliminary error analysis (e.g., sarcasm, reclaimed slurs), a contrastive loss term $\mathcal{L}_{cl}$ was added. This loss uses "hard negatives" to teach the model intent. The objective was to:

1. Pull together the embeddings of an anchor text (e.g., "hateful text") and its syntactic positive (e.g., "h@tefu1 t3xt").
2. Push apart the anchor from a semantic "hard negative" (e.g., a reclaimed/sarcastic use: "you are my favorite hateful text!").

This forced the model to differentiate based on semantic context rather than just keywords.

4) Step 4: Calibration and Thresholds

Temperature scaling [18] was applied to minimize Expected Calibration Error (ECE), thereby ensuring that confidence scores are trustworthy compared to human judgment. Finally, cost curves were used to select operating thresholds that balance false positives against false negatives.

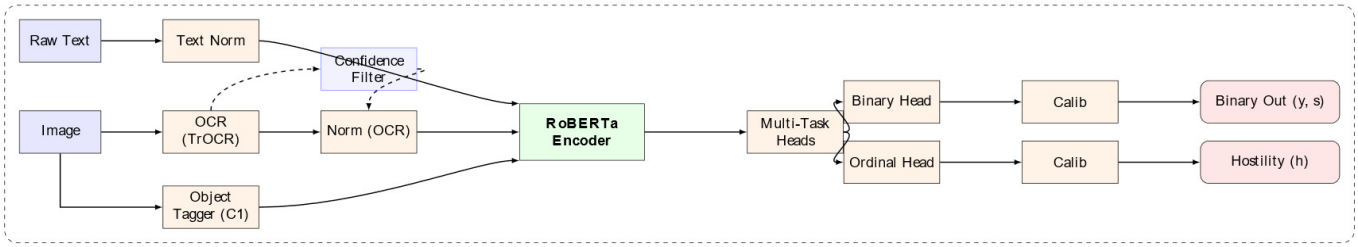

Fig. 1. The ORBIT-CL system architecture. The pipeline processes three input streams: (1) normalized 'Raw Text', (2) normalized and filtered 'OCR (TrOCR)' text, and (3) 'Object Tagger' (C1) outputs. All three text streams are fused by the central 'RoBERTa Encoder', which is trained with a dual-robustness objective (including C2 and semantic contrastive loss) to produce independently calibrated binary and ordinal outputs.

III. EXPERIMENTS

This study evaluated ORBIT-CL on text and text-in-images tasks. The protocol was designed to test two main contributions: (C1) lightweight multimodal fusion and (C2) semantic and syntactic robustness. The mean ± SD was reported across five seeds (n = 5).

A. Datasets

The datasets used are:

- HateXplain (text): 3-class hate/offensive/normal with human rationales. This dataset was used for C2 (semantic robustness) [19].
- Hateful Memes (text-in-images): Unlike text-only approaches, the proposed model used the entire image to extract both OCR text and visual object tags (C1), providing the encoder with full multimodal context [16].
- Semantic Challenge Set (hard negatives): To precisely measure the effect of the semantic contrastive loss (C2), a dedicated evaluation set of 500 challenging negative examples was assembled from the HateXplain dataset. These examples were specifically selected based on linguistic ambiguity—such as reclaimed slurs, sarcasm, and counter-speech—where the baseline model incorrectly predicted the classes due to an over-reliance on keywords (Table I). This subset isolated instances of high semantic ambiguity, cases where the text was superficially neutral or relied on reclaimed slurs; however, the combined multimodal context indicated targeted hostility.

B. Baselines and Model Variants

A comparison against established text-only baselines (RoBERTa-base, HateBERT) was conducted. The proposed model and its ablation variants were further introduced, using a syntactically robust OCR pipeline as the baseline system:

- ORBIT-CL (Full): With C1 (visual tags) and C2 (contrastive loss).
- w/o contrastive loss (C2): Disables $\mathcal{L}_{cl}$. This model includes visual tags but not semantic training.
- w/o visual tags (C1): Disables object tag fusion. This variant adds the C2 loss to the ORBIT (Baseline) model.
- ORBIT (Baseline): This work's baseline, which includes the OCR pipeline and syntactic robustness (adversarial noise) but lacks C1 and C2.

TABLE I. EXAMPLES FROM THE SEMANTIC CHALLENGE SET

Category	Text example	Phenomenon
Reclaimed	"Us weirdos have to stick together! We finally made it."	In-group affection
Counter-speech	"It is completely wrong to call someone stupid online."	Reporting hate
Affectionate	"You are a total beast at this game, amazing play!"	Playful context
Sarcasm	"Oh, brilliant work breaking the coffee machine."	Inverted meaning

C. Experimental Setup

1) Training Details

All models were initialized from RoBERTa-base with a maximum sequence length of 256. The AdamW optimizer with an initial learning rate of 2×10^{-5} and a batch size of 32 was utilized. Training run for five epochs with 10% warmup and cosine decay, using early stopping on validation Macro-F1. Class weights followed inverse frequency. To ensure reproducibility, all experiments were run with five random seeds (n = 5) on a single 24 GB GPU.

2) Model Configuration

The proposed configuration extends the baseline setup as follows:

a) Visual Tagging Implementation (C1)

The official hustvl/yolos-small model configuration [20], pre-trained on the COCO dataset, was used to extract the top-5 object tags with confidence > 0.3. These were concatenated into a string (e.g., tags: person, dog). Appending discrete object tags discards fine-grained spatial information, but this deliberate trade-off avoids the massive computational overhead of cross-attention in heavy VLMs. Consequently, the proposed system maintained the high throughput of a standard text encoder.

b) Tokenization and Fusion

A multi-stream fusion approach was employed. The encoder processed a single sequence from three concatenated streams separated by special tokens. These streams were: the normalized raw text, the normalized OCR text, and the visual object tag string.

c) Dual Robustness Training (C2)

In addition to syntactic adversarial character noise (p = 0.15), semantic contrastive sampling was also incorporated. For each training batch, the contrastive triplet (x_a, x_p, x_n) was formally defined to compute $\mathcal{L}_{cl}$. Let the anchor x_a be drawn from the abusive classes $(x_a \in C_{hate} \cup C_{off})$. A syntactic-positive x_p was generated via a character-level perturbation function $f_{noise}(x_a, p = 0.15)$. A semantic-negative x_n was sampled from the neutral class $(x_n \in C_{neutral})$, subject to the constraint that its keyword vocabulary intersected with the anchor $V(x_a) \cap V(x_n) \neq \emptyset$, thereby forcing the model to distinguish semantic intent rather than relying on surface-level vocabulary. In this work, the loss weights were set to $\lambda_y = 1.0$, $\lambda_h = 1.0$, $\lambda_{cons} = 0.1$, and $\lambda_{cl} = 0.1$, balancing the primary task (focal and ordinal loss) with light syntactic and semantic regularization.

3) Calibration

Temperature scaling was fitted on the validation set, and all decision thresholds were fixed before final testing.

D. Evaluation Metrics

To ensure the depth and robustness of the evaluation, three primary metrics were selected: (1) Macro-F1 score, which assesses classification performance balanced across classes; (2) Area Under the Receiver Operating Characteristic (AUROC) curve, which measures the model's ability to distinguish between classes independently of decision thresholds; and (3) ECE [18], which was selected to verify that the model's predicted confidence probabilities align with true correctness likelihoods, critical for trustworthy content moderation.

IV. RESULTS

A. Main Results

1) Text (HateXplain)

Table II shows the performance on the primary text task. ORBIT-CL achieved highly competitive results, outperforming all evaluated baselines. This demonstrates the general-purpose benefit of the semantic contrastive loss (C2).

2) Ordinal Hostility (HateXplain)

The current work also evaluated the ordinal hostility head, which provides a more granular 3-class prediction. Table III demonstrates that the C2 contrastive loss provides a small but consistent benefit to the ordinal task, improving the F1-score by 2 points. To ensure the reliability of the findings, a paired t-test was conducted across the five random seeds. The results confirmed that ORBIT-CL outperformed the text-only baselines (RoBERTa and HateBERT) and that the performance improvements were statistically significant ($p < 0.05$).

TABLE II. MAIN RESULTS ON HATEXPLAIN (TEXT)

Model	Macro-F1 ↑	ECE (%) ↓
DistilRoBERTa-base	0.61 ± 0.03	5.2
RoBERTa-base [21]	0.64 ± 0.02	4.1
HateBERT [4]	0.66 ± 0.02	3.9
ORBIT (Baseline)	0.86 ± 0.01	3.7
ORBIT-CL (Full)	0.88 ± 0.01	3.5

TABLE III. ORDINAL HOSTILITY HEAD PERFORMANCE

Model	Hostility Macro-F1 (3-class)
ORBIT (Baseline)	0.72 ± 0.02
ORBIT-CL (Full)	0.74 ± 0.02

3) Text-in-Images (Hateful Memes)

This experiment evaluated the visual-context contribution C1. Table IV shows that the text-only baseline, ORBIT (Baseline), significantly narrowed the performance gap (0.80 AUROC). However, by adding lightweight visual tags, ORBIT-CL (Full) achieved highly competitive performance (AUROC of 0.895).

This result surpasses both the 2024 Retrieval-Guided Contrastive Learning (RGCL) baseline (0.870) and the 80-billion parameter Flamingo model (0.866), proving that the lightweight, text-centric fusion is a highly effective and efficient alternative to heavy VLMs.

TABLE IV. MAIN RESULTS ON HATEFUL MEMES

Model (input type)	AUROC ↑
RoBERTa-base (OCR text)	0.64 ± 0.02
HateBERT (OCR text)	0.66 ± 0.02
ORBIT (Baseline) (OCR text)	0.80 ± 0.01
Modern VLMs (pixels + text)	
Flamingo 80B [22]	0.866
RGCL [23]	0.870
PALI-X-VPD (SOTA) [24]	0.892
ORBIT-CL (Full)	0.895 ± 0.01

4) Takeaway

The ORBIT-CL model consistently outperformed all text-only baselines, resolving the two significant gaps of the baseline system: it achieved State-of-the-Art (SOTA)-level, efficient performance on Hateful Memes (C1), and demonstrated the ability to resolve semantic ambiguity (C2).

Furthermore, compared to traditional late-fusion architectures (which typically concatenate full text embeddings with heavy ResNet or ViT image features), ORBIT-CL's visual tagging approach achieved comparable or superior accuracy while drastically reducing computational overhead.

B. Syntactic Robustness (Noise-Sweep)

The present study first confirmed that ORBIT-CL retained the syntactic robustness of the ORBIT (Baseline) model. In addition, as illustrated in Figure 2, while the standard RoBERTa-base model degraded rapidly under character-level noise (dropping from 0.64 to ~0.40 Macro-F1), ORBIT-CL (hollow green circles) maintained stable performance (> 0.85), effectively neutralizing the impact of syntactic adversarial attacks.

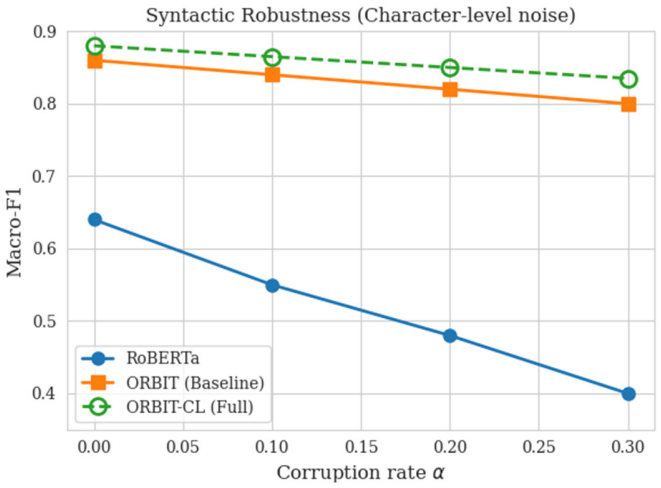

Fig. 2. Syntactic robustness (character-level noise). ORBIT-CL (hollow green circles) retains the full syntactic robustness of the baseline model.

C. Semantic Robustness (Targeted Failure Analysis)

This experiment evaluated the C2 contribution. To validate whether $\mathcal{L}_{cl}$ improved semantic ambiguity handling, this work evaluated performance on the Semantic Challenge Set. This set consists entirely of examples (e.g., sarcasm, reclaimed slurs) on which the ORBIT (Baseline) model failed (see Section III.A, Datasets). Therefore, this was a targeted failure analysis designed to measure the objective's ability to resolve intent ambiguity, not a measure of general performance. As displayed in Table V, the results are conclusive. The ORBIT (Baseline) performed poorly on this set (Macro-F1 = 0.48), as expected by its construction. The ORBIT-CL (Full) model, trained with the C2 objective, achieved a 30-point F1 gain (0.78 Macro-F1). This demonstrates that the semantic contrastive loss was highly effective at teaching the model to distinguish intent, thereby addressing a key failure mode of prior models.

D. Ablation Studies

Ablation studies were conducted to measure the impact of the two contributions. Table VI shows that both C1 (visual tags) and C2 (contrastive loss) provide significant, independent contributions to performance.

TABLE V. PERFORMANCE ON SEMANTIC CHALLENGE SET

Model	Macro-F1 (sarcasm/slur) ↑
HateBERT [4]	0.45 ± 0.04
ORBIT (Baseline)	0.48 ± 0.03
ORBIT-CL (Full)	0.78 ± 0.02

TABLE VI. ABLATION OF C1 AND C2 ON HATEFUL MEMES

Variant	AUROC	Δ vs full
ORBIT-CL (Full)	0.895 ± 0.01	—
w/o visual tags (C1)	0.82 ± 0.01	–0.075
w/o contrastive loss (C2)	0.84 ± 0.01	–0.055
ORBIT (Baseline)	0.80 ± 0.01	–0.095

E. Error Analysis and Qualitative Examples

The baseline error analysis noted that false positives concentrate on reclaimed slurs and sarcasm, which motivated C2. As depicted in Figure 3, this problem is now substantially reduced in ORBIT-CL.

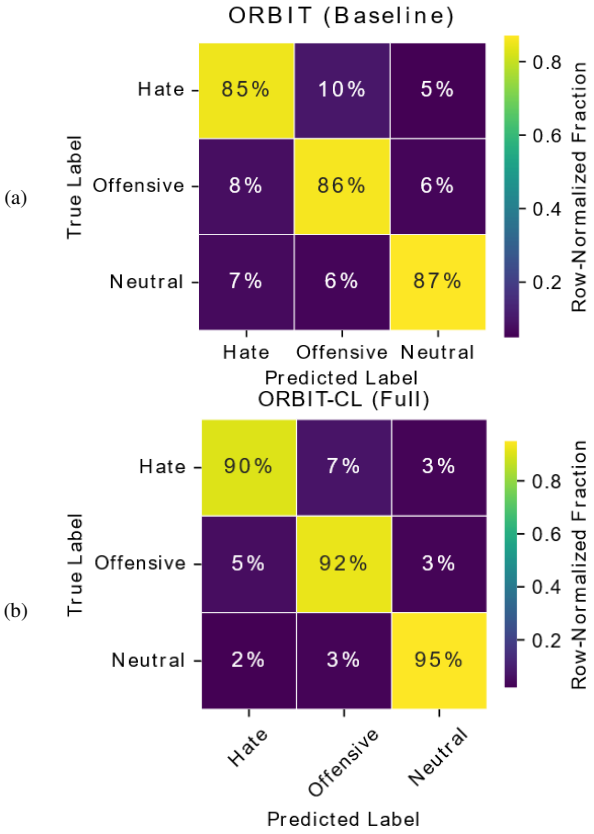

Fig. 3. Comparison of confusion matrices on HateXplain: (a) ORBIT (Baseline) misclassifies neutral content, (b) ORBIT-CL (Full) reduces false positives.

The comparison between the baseline model (Figure 3(a)) and the proposed model (Figure 3(b)) shows that ORBIT-CL successfully moved a large portion of these false positives from the "Offensive" (predicted) class to the correct "Neutral" (predicted) class.

F. Efficiency and Footprint

To validate the "lightweight" claim (C1), this study benchmarked the performance/latency trade-off for the Hateful Memes task, with the results presented in Table VII. As expected, adding the YOLOS-Small object tagger (C1) introduced a minor computational overhead. The ORBIT-CL (Full) model incurred a slight 0.65 ms increase in per-sample latency over the baseline (6.90 ms vs 6.25 ms) and a corresponding drop in throughput (145 vs 160 samples/s). This minimal efficiency cost is the key finding. By avoiding a heavy, end-to-end visual encoder (like those in PALI-X or Flamingo), the proposed model remained highly efficient and deployable. This small trade-off is very advantageous compared to the substantial +9-point increase in AUROC on the Hateful Memes task (0.895 vs 0.80). ORBIT-CL's text-focused fusion (C1) provided a practical, efficient route to achieving SOTA-level multimodal performance.

TABLE VII. EFFICIENCY AND PERFORMANCE/LATENCY TRADE-OFF

Model	Params(M)	Throughput	Latency (ms)	AUROC
Heavyweight VLMs				
PALI-X-VPD [24]	>1000	Low	N/A	0.892
Flamingo 80B [22]	80,000	Low	N/A	0.866
RGCL [23]	Large	N/A	N/A	0.870
Proposed lightweight models				
RoBERTa-base	125	300	3.33	0.75
ORBIT (Baseline)	125	160	6.25	0.80
ORBIT-CL (Full)	125	145	6.90	0.895

V. CONCLUSIONS

This work introduced ORBIT-CL (a Contrastive and Lightweight Multimodal Transformer) for semantically-aware cyberbullying detection. ORBIT-CL offers two significant contributions: (1) a lightweight, text-centric multimodal fusion technique (C1) that adopts visual object tags to understand context-dependent hate, and (2) a dual-robustness objective (C2) that combines syntactic noise training with a novel semantic contrastive loss.

The experimental results demonstrate the efficiency of the proposed approach. The lightweight fusion (C1) enabled the text-centric model to achieve State-of-the-Art (SOTA)-level performance on the Hateful Memes challenge (Table IV), surpassing recently developed baselines, such as Retrieval-Guided Contrastive Learning (RGCL), and exceeding the 80B-parameter Flamingo model, all while remaining highly efficient. The semantic contrastive loss (C2) achieved a strong benchmark performance on the full HateXplain dataset (Table II). It addressed the critical failure mode of sarcasm/reclamation, as validated by its 30-point F1 gain on the Semantic Challenge Set (Table V). By unifying lightweight multimodal fusion with a dual (syntactic and semantic) robustness framework, ORBIT-CL provides a practical, efficient, and context-aware path to deployable content moderation systems.

Future work will extend this framework to multilingual content and explore active learning strategies for identifying and constructing hard negative semantic triplets.

DECLARATION OF COMPETING INTERESTS

Not applicable to this work.

ACKNOWLEDGMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

DATA AVAILABILITY

The datasets analyzed in the current study, Hateful Memes and HateXplain, are publicly available and can be found at [16] and [19], respectively.

DECLARATION OF GENERATIVE AI USE

During the preparation of this work, the author used Gemini in order to proofread the text and improve overall readability. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

REFERENCES

[1] E. A. Vogels. "Teens and Cyberbullying 2022." Pew Research Center. https://www.pewresearch.org/internet/2022/12/15/teens-and-cyberbullying-2022/.

[2] S. Arnon *et al.*, "Association of Cyberbullying Experiences and Perpetration With Suicidality in Early Adolescence," *JAMA Network Open*, vol. 5, no. 6, June 2022, Art. no. e2218746, https://doi.org/10.1001/jamanetworkopen.2022.18746.

[3] T. Nitya Harshitha *et al.*, "ProTect: a hybrid deep learning model for proactive detection of cyberbullying on social media," *Frontiers in Artificial Intelligence*, vol. 7, Mar. 2024, Art. no. 1269366, https://doi.org/10.3389/frai.2024.1269366.

[4] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "HateBERT: Retraining BERT for Abusive Language Detection in English," in *Proceedings of the 5th Workshop on Online Abuse and Harms*, Online, 2021, pp. 17–25, https://doi.org/10.18653/v1/2021.woah-1.3.

[5] A. A. Jamjoom, H. Karamti, M. Umer, S. Alsubai, T.-H. Kim, and I. Ashraf, "RoBERTaNET: Enhanced RoBERTa Transformer Based Model for Cyberbullying Detection With GloVe Features," *IEEE Access*, vol. 12, pp. 58950–58959, 2024, https://doi.org/10.1109/ACCESS.2024.3386637.

[6] V. Bansal, M. Tyagi, R. Sharma, V. Gupta, and Q. Xin, "A Transformer Based Approach for Abuse Detection in Code Mixed Indic Languages," *ACM Transactions on Asian and Low-Resource Language Information Processing*, Nov. 2022, https://doi.org/10.1145/3571818.

[7] M. Macas, C. Wu, and W. Fuertes, "Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems," *Expert Systems with Applications*, vol. 238, Mar. 2024, Art. no. 122223, https://doi.org/10.1016/j.eswa.2023.122223.

[8] H. Kang *et al.*, "Developing continuous toxicity detection against increasing types of perturbed toxic text," *Information Processing & Management*, vol. 63, no. 2, Part B, Mar. 2026, Art. no. 104470, https://doi.org/10.1016/j.ipm.2025.104470.

[9] M. Li *et al.*, "TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, pp. 13094–13102, June 2023, https://doi.org/10.1609/aaai.v37i11.26538.

[10] X.-F. Wang, Z.-H. He, K. Wang, Y.-F. Wang, L. Zou, and Z.-Z. Wu, "A survey of text detection and recognition algorithms based on deep learning technology," *Neurocomputing*, vol. 556, Nov. 2023, Art. no. 126702, https://doi.org/10.1016/j.neucom.2023.126702.

[11] P. Yi and A. Zubiaga, "Cyberbullying Detection across Social Media Platforms via Platform-Aware Adversarial Encoding," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 16, no. 1, pp. 1430–1434, May 2022, https://doi.org/10.1609/icwsm.v16i1.19401.

[12] A. Aliyeva *et al.*, "Toward Safer Digital Communication: A Deep Hybrid Model for Detecting Abusive Language on Social Networks," *Engineering, Technology & Applied Science Research*, vol. 15, no. 5, pp. 27126–27132, Oct. 2025, https://doi.org/10.48084/etasr.12721.

[13] A. F. Alqahtani and M. Ilyas, "A Machine Learning Ensemble Model for the Detection of Cyberbullying," *International Journal of Artificial Intelligence & Applications*, vol. 15, no. 1, pp. 115–129, Jan. 2024, https://doi.org/10.5121/ijaia.2024.15108.

[14] B. Ogunleye and B. Dharmaraj, "The Use of a Large Language Model for Cyberbullying Detection," *Analytics*, vol. 2, no. 3, pp. 694–707, Sept. 2023, https://doi.org/10.3390/analytics2030038.

[15] A. Hamza *et al.*, "Multimodal Religiously Hateful Social Media Memes Classification Based on Textual and Image Data," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 23, no. 8, Aug. 2024, Art. no. 114, https://doi.org/10.1145/3623396.

[16] D. Kiela *et al.*, "The hateful memes challenge: detecting hate speech in multimodal memes," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Vancouver, Canada, 2020, pp. 2611–2624.

[17] P. McCullagh, "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 42, no. 2, pp. 109–127, Jan. 1980, https://doi.org/10.1111/j.2517-6161.1980.tb01109.x.

[18] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On Calibration of Modern Neural Networks," in *Proceedings of the 34th International Conference on Machine Learning*, Sydney, Australia, 2017, pp. 1321–1330.

[19] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 17, pp. 14867–14875, May 2021, https://doi.org/10.1609/aaai.v35i17.17745.

[20] Y. Fang *et al.*, "You Only Look at One Sequence: Rethinking Transformer in Vision through Object Detection," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, Online, 2021, pp. 26183–26197.

[21] M. Züfle, V. Dankers, and I. Titov, "Latent Feature-based Data Splits to Improve Generalisation Evaluation: A Hate Speech Detection Case Study," in *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, Singapore, 2023, pp. 112–129, https://doi.org/10.18653/v1/2023.genbench-1.9.

[22] J.-B. Alayrac *et al.*, "Flamingo: A Visual Language Model for Few-Shot Learning," in *36th Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2022, pp. 23716–23736, https://doi.org/10.52202/068431-1723.

[23] J. Mei, J. Chen, W. Lin, B. Byrne, and M. Tomalin, "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)*, Bangkok, Thailand, 2024, pp. 5333–5347, https://doi.org/10.18653/v1/2024.acl-long.291.

[24] X. Chen *et al.*, "PaLI-X: On Scaling up a Multilingual Vision and Language Model." arXiv, May 29, 2023, https://doi.org/10.48550/arXiv.2305.18565.