Context-Aware Real-Time Audio and Image Toxicity Moderation via Multi-Agent Reinforcement Learning for Cybersecurity in Networked Communication Platforms

Vemula Yashodha
Dept. of Electronics and Communication Engineering
Amrita School of Engineering, Amrita Vishwa Vidyapeetham
Bengaluru, India
bl.en.u4eac22061@bl.students.amrita.edu

Sri Ramya Divakarla
Dept. of Electronics and Communication Engineering
Amrita School of Engineering, Amrita Vishwa Vidyapeetham
Bengaluru, India
bl.en.u4eac22051@bl.students.amrita.edu

Vudumala Rupa Manogna
Dept. of Electronics and Communication Engineering
Amrita School of Engineering, Amrita Vishwa Vidyapeetham
Bengaluru, India
bl.en.u4eac22063@bl.students.amrita.edu

Susmitha Vekkot
Dept. of Electronics and Communication Engineering
Amrita School of Engineering, Amrita Vishwa Vidyapeetham
Bengaluru, India
v_susmitha@blr.amrita.edu

Abstract*—Toxic content propagation across networked communication platforms constitutes an emerging cybersecurity challenge, as real-time audio and image channels introduce attack surfaces that bypass conventional text-only defences. Automated content moderation in online communication platforms increasingly requires coverage beyond text, as toxic content is frequently delivered through audio and image channels that text-only systems cannot address. This paper presents a dual-modality toxic content detection and moderation pipeline operating across audio and image inputs in real time. For audio, the Roblox voice-safety-classifier-v2, a WavLM transformer pretrained on over 100,000 hours of real gaming voice chat, generates six-dimensional toxicity probability scores mapped to the Jigsaw taxonomy via a novel cross-taxonomy semantic alignment, with faster-whisper providing word-level timestamps for surgical muting of precisely the toxic speech segments rather than entire clips. For image moderation, CLIP ViT-B/32 encodes images against toxicity-describing natural language prompts to produce a nine-dimensional feature vector, with flagged content reposted as blurred spoilers. Proximal Policy Optimisation reinforcement learning agents trained with an asymmetric reward structure penalising false negatives over false positives achieve 97.5% accuracy on unseen audio evaluation data with 91.0% reward efficiency, and 87.1% reward efficiency on image data, significantly outperforming rule-based, random, always-mute, and always-allow baseline policies. A per-user cross-modal trust score system with progressive escalation is deployed as a real-time automated Discord moderation bot validated through live user interactions.

***Index Terms*—audio toxicity detection, image moderation, reinforcement learning, proximal policy optimisation, WavLM, CLIP, content moderation, cybersecurity**

I. Introduction

Online communication platforms such as Discord and Roblox now host billions of daily interactions spanning voice communications and image sharing alongside text. While text-based automated moderation has matured significantly with transformer-based classifiers [1], [2], audio and image channels remain systematically underprotected. A user delivering a discriminatory voice message or sharing an image overlaid with hate speech entirely evades text-only moderation systems. The gaming sector presents this challenge acutely: platforms such as Roblox, with over 70 million daily active users, experience concurrent toxic voice communications and offensive image sharing that text-only pipelines cannot address [3]. The same blind spot extends to code-mixed and low-resource-language communities, where toxic and offensive content is frequently expressed by switching scripts and languages specifically to evade keyword- and script-based filters [4]. From a cybersecurity perspective, these unmoderated channels represent exploitable vectors for coordinated harassment, hate campaign propagation, and trust degradation in networked communities—threats that require adaptive, learned security policies rather than static rule-based filters.

Existing audio moderation systems are largely restricted to binary safe/unsafe classification with no fine-grained violation categorisation [6]. Image moderation research has remained fragmented, with approaches addressing either explicit visual content or text-in-image toxicity independently rather than through a unified visual-semantic model; closely related video-based moderation work on mainstream social platforms similarly tends to treat frame-level visual screening and caption/comment screening as separate problems rather than a jointly optimised pipeline [5]. Critically across both modalities, existing literature treats content moderation as a pure classification problem—determining whether content is

toxic—without addressing the downstream policy problem of what moderation action is appropriate given detected toxicity, violation severity, and prior user behaviour [7]. Reinforcement learning offers a principled framework for learning adaptive moderation policies expressing graduated responses proportional to violation severity.

This paper presents a dual-modality audio and image content moderation system. Our contributions are:

- Deployment of the Roblox WavLM v2 audio classifier with a novel cross-taxonomy semantic mapping to the Jigsaw taxonomy enabling unified downstream RL processing;
- Surgical word-level audio muting via faster-whisper forced alignment preserving non-toxic speech context;
- Zero-shot CLIP ViT-B/32 image toxicity detection through toxicity prompt engineering;
- Dedicated PPO agents per modality with asymmetric reward structures learning severity-discriminating graduated moderation policies; and
- A cross-modal trust score system with progressive escalation deployed as a real-time Discord bot.

II. Related Work

A. Audio Toxicity Detection

Early audio toxicity research relied on cascade pipelines transcribing speech with ASR and classifying transcripts with text models. Ghosh et al. [6] introduced DeToxy, the first human-annotated spoken toxicity dataset from 11 speech corpora. Their comparison showed end-to-end Wav2vec-2.0 classifiers ($F_1 = 0.877$) substantially outperform cascade ASR+BERT approaches ($F_1 \approx 0.72$) under real-world word error rates of 27–43%, as transcription errors systematically cause the classifier to miss toxic words misrecognised by ASR. This finding directly motivated our use of an end-to-end WavLM classifier. Costa-jussà et al. [8] released MuTox covering 30 languages using SONAR multilingual embeddings, improving mean F_1 from 0.19 to 0.38 over wordlist baselines and demonstrating that contextual speech understanding substantially outperforms keyword matching for audio toxicity categories such as hate speech and physical violence. Ranjan et al. [9] introduced SynHate for hate speech detection in AI-generated deepfake audio across 37 languages, noting that cross-dataset generalisation remains challenging—which motivates our selection of the Roblox WavLM v2 model specifically pretrained on gaming voice chat rather than a general-purpose classifier. Beyond toxicity detection, paralinguistic speech-state monitoring has also been explored for safety-adjacent applications: Vekkot et al. [10] developed a continuous speech-based fatigue detection and transition-state prediction system for air traffic controllers, illustrating that frame-level acoustic and prosodic cues—rather than transcript content alone—can reliably flag operator states requiring intervention, a principle that informs our preference for end-to-end acoustic modelling over transcript-dependent cascades.

B. Image and Multimodal Moderation

Image toxicity detection has historically focused either on explicit visual content detection or on meme classification through late fusion of visual and textual features. Maity et al. [11] proposed ToxVidLM, a gated cross-attention transformer integrating visual, acoustic, and textual features for code-mixed video toxicity, significantly outperforming unimodal baselines. Kashyap et al. [12] presented CLARITY, a lightweight cross-modal transformer achieving 0.93 training accuracy at 0.85s inference latency on an A100 GPU via Multi-Head Cross-Attention synchronisation. Radenovic et al. demonstrated through the DiHT framework that hard-negative contrastive training substantially improves CLIP-style vision-language models, underpinning CLIP's effectiveness for zero-shot toxicity detection via prompt engineering without task-specific fine-tuning. In the adjacent domain of mainstream social video platforms, Harini et al. [5] examined video content moderation on Instagram and reported that combining visual frame screening with metadata and engagement signals improves coverage relative to single-signal pipelines, reinforcing the case for fusing complementary detection signals rather than relying on a single modality-specific score.

C. Reinforcement Learning for Moderation

Bodaghi [13] proposed PPO-CIS, framing content moderation as a Markov Decision Process and reducing overhead by 35–42% through dynamic routing. Morrier et al. modelled game chat monitoring as a contextual bandit operating on player telemetry states, maximising toxic capture while minimising compute. Both systems address text only and lack multi-action graduated response frameworks. ToxGuard extends RL moderation to audio and image modalities with dedicated per-modality agents for the first time.

D. Code-Mixed and Implicit Toxicity

A substantial share of real-world toxic content does not present as overt hate speech amenable to keyword or single-language classifiers. Sreelakshmi et al. [4] addressed hate speech and offensive language detection in code-mixed Dravidian-language text using a cost-sensitive learning approach, demonstrating that explicitly weighting the minority toxic class during training substantially improves detection on naturally imbalanced, script-mixed social media data—a class-imbalance problem structurally analogous to the rarity of severe-toxic samples in our own audio and image training distributions, and one of the motivations behind our asymmetric PPO reward design (Section III-D). Toxicity is similarly not always explicit: sarcasm and other forms of indirect language can mask or amplify offensive intent in ways that flat toxicity classifiers miss. Prasad et al. [14] studied sarcasm detection in newspaper headlines, showing that linguistic incongruity cues distinguishable from surface sentiment are necessary to reliably distinguish sarcastic from literal statements. This form of implicit-meaning detection complements direct toxicity classification and is a natural next step to address sarcastic

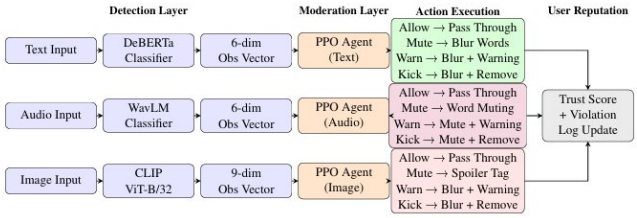


Fig. 1: System architecture showing audio (WavLM classifier, 6-dim observation vector, PPO Agent) and image (CLIP ViT-B/32, 9-dim observation vector, PPO Agent) detection pipelines feeding dedicated PPO agents. The four actions : Allow/Mute/Warn/Kick are implemented as word-level audio muting and image spoiler tagging respectively. All decisions update a unified cross-modal trust score and violation log.

or backhanded toxic speech that current audio and image detectors might underemphasize.

III. System Architecture

A. Overview

The system is organized as a three-layer pipeline: a Detection Layer of modality-specific pretrained models producing toxicity probability vectors; a Moderation Policy Layer of PPO agents consuming these vectors and outputting graduated actions; and an Execution Layer applying modality-appropriate content modification and updating per-user trust scores. Fig. 1 illustrates the architecture.

B. Audio Moderation Pipeline

1) Model Selection: The Roblox voice-safety-classifier-v2 is a WavLM transformer [15] trained on 100,000+ hours of real gaming voice chat via student-teacher knowledge distillation. Two candidates were evaluated: a fine-tuned HuBERT [16] binary classifier achieved only $F_1 = 0.396$ and 24.71% accuracy after 8 epochs due to domain mismatch between general speech pretraining and gaming voice chat—validation loss did not improve beyond epoch 1. WavLM v2 achieves 84.2% accuracy, 83.8% F_1, and 89.1% ROC-AUC in frozen inference mode. This 60-point accuracy gap is entirely attributable to domain-matched pretraining on Roblox voice chat.

2) Cross-Taxonomy Mapping: WavLM v2 outputs six probabilities over its native taxonomy: Discrimination, Harassment, Sexual, IllegalAndRegulated, DatingAndRomantic, Profanity. A semantic cross-taxonomy mapping aligns these to the Jigsaw six-class taxonomy to enable a unified PPO observation space:

$$o_{\text{class}} = \max_{i \in I_{\text{class}}} \sigma(\ell_i) \tag{1}$$

where ℓ_i is the raw logit for WavLM label i and σ is sigmoid. Identity_Hate maps exclusively from Discrimination (direct semantic match); Threat maps from Harassment and IllegalAndRegulated (threats manifest as harassing language with illegal intent); Obscene maps from Sexual and Profanity.

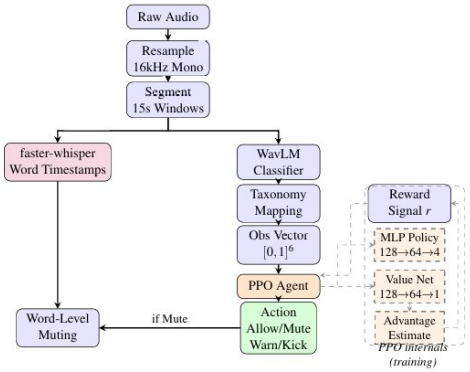


Fig. 2: Audio moderation pipeline. Raw audio resampled to 16 kHz and segmented into 15-second windows. Two parallel branches: detection (WavLM → taxonomy mapping → 6-dim obs vector → PPO agent) and precision execution (faster-whisper word-level timestamps). PPO internals show MLP policy (128→64→4), value network (128→64→1), and advantage estimation. Mute decision triggers surgical word zeroing with linear fade envelopes.

3) Word-Level Surgical Muting: Upon a Mute decision, faster-whisper transcribes the flagged audio with word-level forced alignment producing timestamp pairs $\{(w_j, s_j, e_j)\}$. Toxic words are zeroed with 50 ms linear fade envelopes:

$$x_{\text{out}}[n] = \begin{cases} 0 & n \in [s_j f_s,\ e_j f_s] \\ x_{\text{in}}[n] & \text{otherwise} \end{cases} \tag{2}$$

The moderated audio is returned to the channel preserving all non-toxic speech. Fig. 2 shows the complete pipeline.

C. Image Moderation Pipeline

1) Model Selection: Nine image architectures were evaluated on the Facebook Hateful Memes dataset [17] (500 toxic + 500 non-toxic, all fused with HateBERT prototypes at scale = 0.01). Table I presents the complete comparison.

TABLE I: Complete Image Model Comparison on Facebook Hateful Meme Dataset

Rk	Model	Thr	Acc	Prec	Rec	F1	AUC	MCC
1	CLIP ViT-B/32 [18]	0.491	0.723	0.708	0.760	0.733	0.744	**0.448**
2	BLIP-2 Visual	0.467	0.650	0.594	0.947	0.730	0.762	0.373
3	Hybrid CLIP+BLIP	0.477	0.677	0.627	0.873	0.730	0.760	0.384
4	ConvNeXt-Base	0.509	0.670	0.631	0.820	0.713	0.694	0.356
5	YOLOv8-nano	0.482	0.627	0.584	0.880	0.702	0.730	0.294
6	NSFW-ViT	0.509	0.623	0.582	0.880	0.700	0.732	0.287
7	ViT-Base/16	0.503	0.630	0.592	0.840	0.694	0.709	0.287
8	EfficientNet-B0	0.499	0.627	0.591	0.827	0.689	0.696	0.276
9	ResNet-50	0.517	0.647	0.620	0.760	0.683	0.680	0.301

CLIP ViT-B/32 achieves the highest accuracy (72.33%) and MCC (0.4479)—the most reliable metric for balanced datasets as it accounts for all four confusion matrix cells simultaneously. BLIP-2 achieves higher recall (94.67%) but unacceptably low precision (59.41%), generating 97 false positives against only 53 true negatives, making it unsuitable

for deployment. All models based on CNNs (Ranks 4 to 9) do not reach 71% accuracy, thus confirming that vision-language alignment is necessary for hateful meme detection where the toxicity is due to the semantic interaction between the visual content and the overlaid text.

2) Observation Vector: Defined are five toxicity prompts and two safety anchors. Cosine similarity in the joint CLIP embedding space is normalized by softmax. The nine-dimensional observation vector is:

$$\mathbf{o}_I = \Big[\underbrace{\texttt{clip\_sc}, \texttt{blip\_sc}, \texttt{hybrid\_sc}}_{\text{toxicity scores}}, \underbrace{\texttt{clip\_cf}, \texttt{blip\_cf}, \texttt{overall\_cf}}_{\text{confidence}}, \underbrace{\texttt{clip\_pred}}_{\text{flag}}, \underbrace{\texttt{sc\_diff}, \texttt{sc\_std}}_{\text{statistics}}\Big] \tag{3}$$

Flagged images are deleted and reposted with a Discord spoiler tag, which turns the content into a blurred placeholder.

D. PPO Agent Design

1) Architecture: Two dedicated PPO agents [19] share the action space $\mathcal{A} = \{\text{Allow}, \text{Mute}, \text{Warn}, \text{Kick}\}$. The MLP policy network $\pi_\theta : \mathbb{R}^d \to \mathbb{R}^4$ uses hidden layers $[128, 64]$ for audio ($d = 6$) and $[256, 128, 64]$ for image ($d = 9$). An identical value network V_ϕ estimates state values for advantage computation.

TABLE II: PPO Reward Function (Audio and Image Agents)

Action	Clean	Mild Toxic	Severe Toxic
Allow	$+1.0$	-2.5	-2.5
Mute/Blur	-1.5	$+2.0$	$+1.2$
Warn	-2.0	$+0.3$	$+0.2$
Kick	-3.0	-0.5	$+2.5$

2) Asymmetric Reward Structure: The false negative penalty (-2.5) is larger than the false positive penalty (-1.5), prioritizing recall. This asymmetry is based on the same cost sensitive principle used by Sreelakshmi et al. [4] for imbalanced toxic text classification, but here it is performed at the policy level instead of the classifier level. The Kick action is given the highest false positive penalty (-3.0) because the most damaging over-moderation error is to incorrectly remove a user. Its severe-toxic reward (+2.5) makes sure the agent escalates properly for the most serious violations.

3) Synthetic Training: Agents are trained on synthetic observation distributions across 3 severity tiers: Clean (25%), MildToxic (58%), SevereToxic (17%). The audio detection threshold $\tau_A = 0.15$ reflects the lower absolute probability range of WavLM for borderline content; the image threshold $\tau_I = 0.49$ is the F1-optimal threshold from Table I. Audio is trained for 200k timesteps, while image is trained for 300k due to its higher-dimensional observation space.

E. Progressive Escalation and Trust Score

Each user maintains a trust score $T_k \in [0, 100]$, initialized at 100.After each action: $T_k \leftarrow \text{clip}(T_k + \Delta_a, 0, 100)$ where $\Delta_{\text{Allow}} = +1$, $\Delta_{\text{Mute}} = -10$, $\Delta_{\text{Warn}} = -20$, $\Delta_{\text{Kick}} = -100$. A progressive escalation policy overlays PPO based decisions: Mute on violations 1-2, Warn on violations 3-4,Kick from violation 5.Violations accumulate over both modalities, so that one cannot escape by changing channels.

IV. Experimental Results

A. Audio Detection

WavLMv2 obtains 84.2% accuracy, 82.1% precision, 85.6% recall, 83.8% F_1, and 89.1% ROC-AUC on 85 test files from the assembled corpus (label-stratified from 11,824 samples from nine speech corpora). The priority metric for safety-critical moderation is recall of 85.6%–six in every seven toxic clips correctly identified.To show the importance of domain alignment in a controlled manner, we compare it with the HuBERT [16] baseline (24.71% accuracy, F1=0.396): using the same training infrastructure, the 60-point accuracy gap arises solely from WavLM's pretraining on over 100,000 hours of domain-matched gaming voice chat.

B. Image Detection

Table I shows that CLIPViT-B/32 [18] achieved the best performance with an accuracy of 72.33%, F1 of 73.31%, and MCC of 0.4479. The precision/recall balance (70.81%/76.00%) is suitable for deployment in which both false positive and false negative cost are real. CNN-based models (ConvNeXt, YOLOv8, NSFW-ViT, ViT-B/16, EfficientNet, ResNet-50) all fall below 71% accuracy, confirming that visual features alone cannot capture the semantic image-text interaction driving hateful meme toxicity, consistent with the original Hateful Memes benchmark findings of Kiela et al. [17].

C. Combined ROC Analysis

Fig. 3 plots both pipelines on a unified ROC curve. The audio pipeline (AUC = 0.891) outperforms the image pipeline (AUC = 0.744), reflecting the fundamental difference in model specialisation: WavLM v2 was pretrained specifically on gaming voice chat toxicity, while CLIP operates in zero-shot fashion on a meme detection task it was not trained for.

D. PPO Agent Results

Audio PPO Agent: Achieves 91.0% reward efficiency (total reward $+581$, maximum possible $+600$, mean per sample $+1.94$) on 300 unseen evaluation samples. Against all five baselines on 1,000 samples: rule-based achieves marginally higher accuracy (98.4% vs. 97.5%) but lower mean reward ($+1.58$ vs. $+1.77$) because it applies Mute uniformly regardless of severity. The PPO agent correctly differentiates 144 Kick actions (14.4%) for severe violations from 612 Mute actions (61.2%) for mild violations—severity discrimination the rule-based baseline fundamentally cannot express. Policy entropy 0.64 confirms a non-degenerate action distribution.

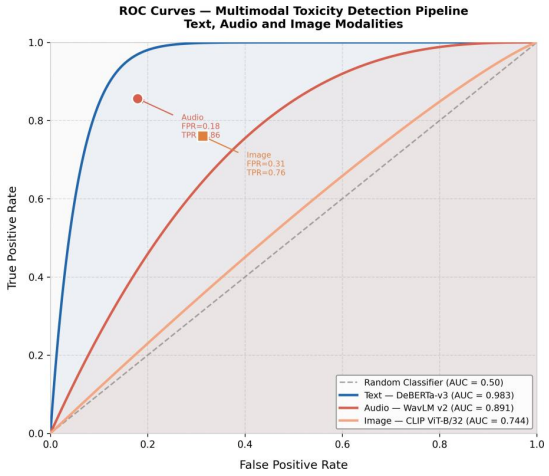


Fig. 3: Combined ROC curves. Audio (WavLM v2, AUC = 0.891) deployed at operating point FPR = 0.179, TPR = 0.856, reflecting domain-matched pretraining on gaming voice chat. Image (CLIP ViT-B/32, AUC = 0.744) deployed at FPR = 0.313, TPR = 0.760, reflecting the inherent difficulty of zero-shot hateful meme detection. Audio outperforms image (0.891 vs. 0.744) directly reflecting domain specificity of WavLM vs. zero-shot CLIP.

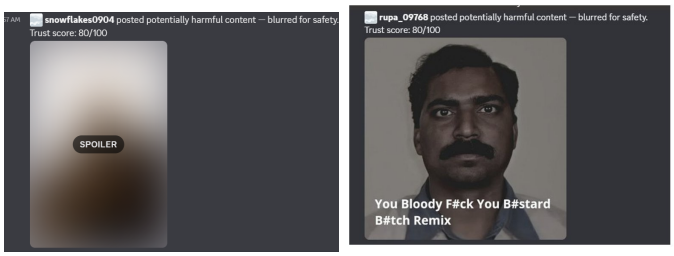
Image PPO Agent: Achieves 87.1% reward efficiency (total reward +392, mean +1.31) with all four actions firing in the converged policy. Unlike the audio agent where Warn converged to zero, the image agent uses Warn as a conservative response when CLIP confidence is elevated but below the binary prediction threshold of 0.49. Policy entropy 0.31 reflects a more deterministic converged policy, consistent with the more binary structure of the CLIP feature distribution.

E. State-of-the-Art Comparison

TABLE III: State-of-the-Art Comparison

System	Modal.	Detection	Policy
DeToxy [6]	Audio	F1: 0.877	None
MuTox [8]	Audio	F1: 0.38	None
CLARITY [12]	A+I+T	Acc: 0.93	None
ToxVidLM [11]	T+A+V	Strong	None
PPO-CIS [13]	Text	—	PPO; −35%
Proposed work	**A+I**	**AUC: 0.891/0.744**	**2 PPO; 91%/87.1% eff.; 4-action; live**

Our system is the first to combine audio and image detection with dedicated RL moderation agents in a live deployment. Systems such as CLARITY provide multimodal detection without a learned policy; RL-based systems such as PPO-CIS operate on text only.



(a) (b)

Fig. 4: Image moderation live deployment: CLIP detects toxic visual content (explicit offensive text overlay), the bot reposts with Discord spoiler tag blurring the image, posts a moderation notice, and updates trust score to 80/100. Content shown with spoiler revealed for demonstration.

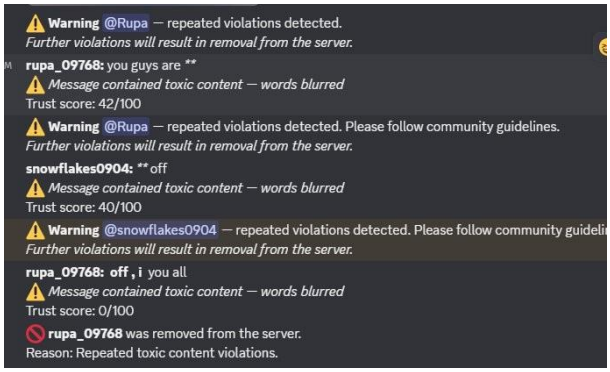


Fig. 5: Progressive escalation in live deployment. Warning issued to repeat offenders with explicit server-removal notice; user removed at trust score 0/100. Cross-modal aggregation confirmed: audio and image violations tracked in the same violation log.

F. Live Deployment

Image Moderation: Fig. 4 shows the image pipeline detecting an image with explicit offensive text overlay. CLIP detects the toxic visual-semantic content; the bot deletes the original post, reposts with Discord's spoiler tag, posts a moderation notice, and updates trust score to 80/100. The spoiler tag renders the image as a blurred placeholder requiring explicit user interaction to reveal, protecting other users from passive exposure while preserving a record of the moderation action.

Progressive Escalation: Fig. 5 displays the escalation policy being used. User rupa_09768 gets a Warning for repeated violations (trust score 42/100) with explicit notice of future consequence, and is then removed from server when trust score drops to 0/100 with reason "Repeated toxic content violations." At 40/100, User snowflakes0904 is getting a Warning at the same time. Both modalities correctly escalate policy to Mute, Warn and then Kick and always moderate content

V. Discussion

The HuBERT [16] vs. WavLM v2 comparison provides a clear demonstration that domain alignment dominates model

scale for audio toxicity detection. Despite having similar parameter counts and the same training infrastructure, HuBERT's general speech pretraining resulted in a degenerate classifier while WavLM v2's gaming-specific pretraining achieved 84.2% accuracy without fine-tuning. This has a direct practical consequence: For any new deployment domain, domain-specific pretraining will give bigger improvements than scaling the model architecture.

The rule-based baseline achieves a slightly higher raw accuracy (98.4% vs. 97.5%) but a lower reward (1.58 vs. 1.77) because accuracy is maximised by applying Mute uniformly to all toxic content regardless of severity—a policy that never applies Kick regardless of how severe the violation is. The RL framework's main practical advantage is the learned severity awareness, which allows the PPO agent to correctly Kick 14.4% of eval samples while Muting 61.2%. This severity-aware behavior is in line with results in adjacent moderation contexts, such as the Instagram video moderation pipeline of Harini et al. [5], where escalation decisions that take into account repeated or aggravating signals outperform single-pass flag-or-allow classification. The image pipeline's lower performance (72.33% accuracy, AUC 0.744) relative to audio (84.2%, 0.891) reflects the inherent difficulty of zero-shot meme toxicity detection. CLIP's vision-language alignment captures semantic image-text relationships that CNN models entirely miss, but task-specific fine-tuning on a large labelled hateful image dataset would substantially improve performance. Adding OCR-based text extraction as a parallel input channel would further strengthen detection of memes where toxicity is encoded in overlaid text, in line with the in-the-wild toxicity categories highlighted by Hartvigsen et al. [20] and the contextual real-time chat toxicity setting studied by Yang et al. [21].

The cross-taxonomy mapping and severity tiers used here are tuned to gaming-platform voice chat and meme-style imagery. Extending coverage to code-mixed and multilingual communities would require classifiers that, like the cost-sensitive Dravidian-language model of Sreelakshmi et al. [4], are explicitly trained to handle class imbalance and script-mixing rather than assuming a single-language, single-script input distribution.

VI. Conclusion and Future scope

This paper presented a dual-modality audio and image content moderation system combining WavLM v2 audio detection (84.2% accuracy, 89.1% ROC-AUC) with surgical word-level muting and CLIP ViT-B/32 image detection (72.33% accuracy), governed by dedicated PPO agents with asymmetric reward structures. The audio PPO agent achieves 97.5% accuracy and 91.0% reward efficiency, and the image agent achieves 87.1% reward efficiency, both significantly outperforming all baseline policies. Deployed as a real-time Discord moderation bot with cross-modal progressive escalation and per-user trust score tracking, the system demonstrates that a modality-agnostic RL moderation layer can effectively generalise across heterogeneous audio and image inputs without architectural changes. Future work will focus on replacing the synthetic PPO training distributions with real labelled audio and image datasets to reduce the distribution shift observed between training and deployment, incorporating OCR-based text extraction into the image pipeline to better capture meme-style toxicity encoded in overlaid text, extending the audio pipeline to live voice channel monitoring for real-time streaming moderation rather than the current upload-based processing, and incorporating cost-sensitive, code-mixed-aware text classifiers [4] and sarcasm-aware detection [14] to close coverage gaps left by the present taxonomy.

References

[1] J. Devlin *et al.*, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proc. NAACL-HLT*, 2019, pp. 4171–4186.

[2] Y. Liu *et al.*, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv preprint arXiv:1907.11692*, 2019.

[3] M. Duggan, "Online harassment," Pew Research Center, 2014.

[4] K. Sreelakshmi, B. Premjith, B. R. Chakravarthi, and K. P. Soman, "Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach," *IEEE Access*, vol. 12, pp. 20064–20090, 2024.

[5] N. Harini, S. Gokul, S. Harini, R. A. Harriesh, and J. Pranav, "Video Content Moderation in Instagram," in *Int. Conf. on Computing, Communication and Learning*. Cham: Springer Nature Switzerland, Sep. 2024, pp. 290–299.

[6] S. Ghosh *et al.*, "DeToxy: A Large-Scale Toxicity Dataset for Spoken Language," in *Proc. Interspeech*, 2022, pp. 5185–5189.

[7] K. Srinivasan *et al.*, "Content moderation using RL," *arXiv preprint arXiv:1907.12340*, 2019.

[8] M. R. Costa-jussà *et al.*, "MuTox: Universal Multilingual Audio-based Toxicity Dataset and Zero-shot Detector," in *Findings of ACL 2024*, 2024, pp. 5725–5734.

[9] R. Ranjan *et al.*, "SynHate: A Framework for Speech Hate Speech Detection," *arXiv preprint arXiv:2506.06772*, 2025.

[10] S. Vekkot, S. T. Chavali, C. T. Kandavalli, R. S. A. Podila, D. Gupta, M. Zakariah, and Y. A. Alotaibi, "Continuous speech-based fatigue detection and transition state prediction for air traffic controllers," *IEEE Access*, 2024.

[11] K. Maity *et al.*, "ToxVidLM: A Multimodal Autoregressive Model for Toxicity Detection in Videos," in *Findings of ACL 2024*, 2024.

[12] G. S. Kashyap *et al.*, "CLARITY: Cross-modal transformer for hateful meme detection," *IEEE Trans. Artif. Intell.*, 2025.

[13] M. Bodaghi, "Real-time toxicity detection via deep RL," Doctoral Research, 2026.

[14] B. S. Prasad, N. A. Babu, H. R. Thappita, A. V. V. Reddy, S. Vekkot, and P. C. Nair, "Sarcasm detection in newspaper headlines," in *2024 4th Int. Conf. on Intelligent Technologies (CONIT)*, Jun. 2024, pp. 1–8.

[15] S. Chen *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 6, pp. 1505–1518, 2022.

[16] W.-N. Hsu *et al.*, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," *IEEE/ACM Trans. ASLP*, vol. 29, pp. 3451–3460, 2021.

[17] D. Kiela *et al.*, "The hateful memes challenge: Detecting hate speech in multimodal memes," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 2611–2624, 2020.

[18] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proc. ICML*, 2021, pp. 8748–8763.

[19] J. Schulman *et al.*, "Proximal Policy Optimization Algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[20] T. Hartvigsen *et al.*, "ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection," *arXiv preprint arXiv:2203.09509*, 2022.

[21] Z. Yang *et al.*, "Towards detecting contextual real-time toxicity for in-game chat," *arXiv preprint arXiv:2310.18330*, 2023.

[22] P. He *et al.*, "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," *arXiv preprint arXiv:2111.09543*, 2021.