Soham S. Jadhav, 2026, 14:3
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Beyond Text: A Comprehensive Survey of Multimodal Content Moderation Architectures in Enterprise Environments

Soham S. Jadhav, Omkar N. Gadakh, Atharv S. Gaikwad, Nisha D. Patil
Dept. Artificial Intelligence and Data Science
MET's Institute of Engineering, Nashik, India

Abstract- Modern enterprise workflows have evolved beyond plain text, now heavily relying on a mix of audio, video, and image-based exchanges. Consequently, the traditional paradigms of content moderation are becoming obsolete. Traditional moderation tools often fail in corporate settings because they were built for public platforms, prioritizing retroactive cleanup rather than real-time privacy and prevention, and relying on static, modality specific classifiers that fail to meet the real-time, privacy-centric, and context-aware demands of modern enterprise environments. This paper presents a comprehensive review of the state-of-the art in intelligent content moderation, analysing over 15 distinct methodologies ranging from Large Language Model (LLM)based guardrails to multimodal fusion architectures. We critically examine the transition from rigid API-based moderation to dynamic "Policy-as-Prompt" frameworks, evaluating their efficacy in handling code-mixed languages, audio-visual semantics, and organizational

Keywords— Multimodal Moderation, Enterprise Security, Large Language Models (LLMs), Policy-as-Prompt, Gatekeeper Architecture, Audio-Visual Fusion.

I. INTRODUCTION

The proliferation of collaborative digital platforms (e.g., Slack, Microsoft Teams, Discord) has fundamentally altered the landscape of organizational communication. Unlike public social media, where the primary moderation goal is to filter toxicity or hate speech on a global scale, organizational environments require nuanced enforcement of specific compliance rules, data leakage prevention (DLP), and professional conduct standards [1].

However, a significant dichotomy exists in current research. While significant strides have been made in text-based safety exemplified by systems like Llama Guard [1] and LlamaFirewall [2] enterprise communication is increasingly multimodal. Employees share screenshots containing sensitive data, send voice notes for quick updates, and utilize memes that may carry implicit bias. Current unimodal systems fail to capture the semantic interplay between these data types. For instance, a benign image coupled with a sarcastic voice note may constitute harassment, a nuance lost on disjointed classifiers[4].

This review paper surveys the architectural shift towards **Unified Multimodal Moderation**. We analyse the limitations of current "reactive" systems that flag content only after it appears in the chat stream, contrasting them with emerging "proactive" or "Gatekeeper" architectures that intercept and validate content pre-delivery. By synthesizing literature from 2019 to 2025, we map the trajectory of moderation technologies and highlight the specific challenges of latency, adaptability, and privacy in closed organizational networks.

II. BACKGROUND AND TAXONOMY

To structure this review, we define three core concepts that have emerged as pillars of modern moderation research.

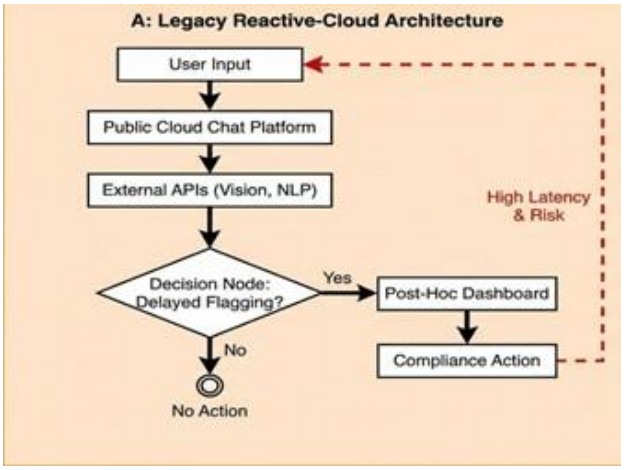
The Policy-as-Prompt Paradigm

Traditional moderation relied on training supervised classifiers on labelled datasets (e.g., "Hate" vs. "Not Hate"). This approach is brittle; changing a company policy often requires retraining the model. The **Policy-as-Prompt** paradigm, discussed extensively by Palla et al. [3], leverages the zero shot capabilities of Large Language Models (LLMs). In this framework, policies are encoded as natural language prompts (e.g., "Block financial disclosures" or "Flag harassment"), allowing organizations to update rules dynamically without incurring technical debt or retraining costs.

Multimodal Fusion Approaches

Fusion refers to the strategy of combining different modalities such as text, audio, and visual data into a coherent representation for analysis. Guo et al. [6] categorize these strategies into three primary types:

- **Early-Stage Integration:** Merging input data at the raw feature level prior to analysis.
- **Late-Stage Aggregation**: Analysing each data type (e.g., text, video) independently and then synthesizing the final outputs.
- **Unified Abstraction:** A nascent approach where all modalities are converted into a single semantic space (typically text) for unified processing by a Reasoning LLM.



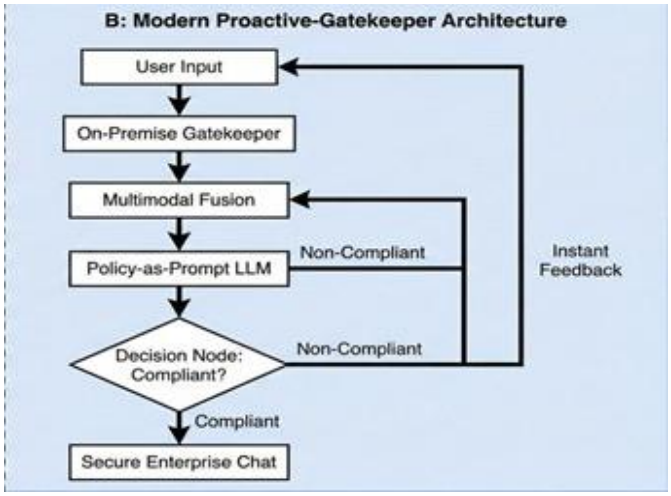


Fig. 1. The evolution of moderation architectures in enterprise environments.

The Gatekeeper Architecture

Unlike *Post-Hoc* moderation, which functions as a cleanup mechanism by flagging content after it has been posted, a *Gatekeeper* architecture is designed to sit between the user input and the broadcast channel. It acts as a real-time validator, intercepting messages before delivery. This architecture imposes strict latency constraints, typically requiring sub-3second processing times to maintain the flow of conversation.

III. LITERATURE REVIEW

Our analysis of the literature reveals four distinct clusters of research focus: LLM-Based Safety, Multimodal Benchmarking, Audio-Semantic Analysis, and Human-Centric Moderation.

LLM-Based Guardrails and Reasoning

The deployment of Large Language Models (LLMs) for safety has seen rapid adoption. Inan et al. [1] introduced **Llama Guard,** a 7B parameter model instruction-tuned on a safety taxonomy. It demonstrated that LLMs could effectively replace rigid APIs, offering adaptability to new rule sets through zero-shot prompting. Building on this, Chennabasappa et al. [2] proposed LlamaFirewall, which focuses on securing AI agents from adversarial inputs. While ***LlamaFirewall*** is highly effective at stopping adversarial prompt attacks (achieving less

than 2% success rates), it is less optimized for the nuanced interpersonal conflicts common in team chats.

Multimodal Benchmarks and Detection

TABLE 1: Comparative summary of critical research gaps in existing moderation systems

Ref / System	Critical Research Gap
Llama Guard [1]	Unimodal (text-only) moderation; lacks organizational context awareness and enterprise-level adaptability.
LlamaFirewall [2]	Focuses on AI agent protection rather than human group communication; limited support for multimodal inputs.
Policy-as-Prompt [3]	High sensitivity to prompt formulation; not evaluated on real-time multimodal communication streams.
MM-SOC [5]	Restricted to offline benchmarking; reveals significant modality gaps in zero-shot evaluation scenarios.
HateMM [8]	Computationally expensive and struggles with implicit semantics such as memes and indirect expressions.
MSCMGTB [7]	Scalability limitations; lacks audio processing and fails to meet real-time latency requirements (<3 seconds).
Semantic Speech [13]	Unimodal audio analysis; unable to correlate speech signals with textual or visual context.
Multimodal XAI [4]	Primarily theoretical; lacks practical implementation for high-speed, real-time moderation systems.
Proposed Framework	Addresses the above gaps through unified multimodal processing and proactive pre-delivery policy enforcement.

The transition to multimodal moderation is currently hindered by a lack of diverse benchmarks. Jin et al. [5] introduced **MM-SOC**, a benchmark suite for social media tasks involving text, images, and video. Their findings highlight a significant "modality gap" multimodal LLMs often perform significantly worse than specialized unimodal models without extensive fine-tuning.

Koushik et al. [8] demonstrated that video moderation becomes significantly more reliable when audio is transcribed into text first, reporting notable accuracy gains compared to processing raw video frames alone.

Audio and Speech Analysis

Audio moderation remains the least developed modality in enterprise tools. Tyagi and Szen´ asi [10] reviewed semantic´ speech analysis, noting that while hybrid models (combining CNNs and LSTMs) are effective for emotion recognition, they often lack the contextual awareness required to detect specific policy violations. However, recent advances in robust speech recognition, such as **Whisper** proposed by Radford et al. [11], have enabled high-fidelity transcription (Word Error Rate ≈ 3%) even in noisy environments. This suggests that the most viable architectural path for audio moderation is **Transcribe then-Moderate.**

Human Factors and Explainability

The introduction of AI moderators fundamentally alters group dynamics. Yu et al. [9] conducted a systematic review of AI-assisted moderation, finding that opacity in automated decisions leads to user distrust. This aligns with the call for **Explainable AI (XAI**) in moderation systems. Users in a professional setting need to know *why* a message was blocked (e.g., "Violated NDA Policy") rather than receiving a generic error [4].

IV. COMPARATIVE ANALYSIS OF ARCHITECTURES

Based on the surveyed literature, we classify current moderation systems into two primary architectures: **Reactive-Cloud and Proactive-On-Premise.**

The **Reactive-Cloud** model, employed by standard social platforms, relies on asynchronous API calls to external services

(e.g., OpenAI or Google Cloud Vision) after a message has been posted. While this approach scales well for public social media, it fails the privacy and immediacy tests required for enterprise use cases. Data must leave the secure perimeter, and enforcement often occurs seconds or minutes after the violation.

In contrast, the **Proactive-Gatekeeper** model, while computationally heavier, ensures that policy-violating content never permeates the organization's digital memory. It utilizes localized or private-cloud inference to validate content *pre-delivery*.

Table I summarizes the critical research gaps identified across the reviewed literature, highlighting the specific deficiencies that emerging gatekeeper architectures aim to address. The analysis indicates a clear trajectory in the field: a move away from disjointed, unimodal classifiers towards unified, reasoning-based architectures.

V. CHALLENGES AND RESEARCH GAPS

Despite the architectural advances discussed, several critical gaps impede the widespread adoption of comprehensive multimodal moderation in enterprise environments.

The Latency-Accuracy Trade-off

Real-time organizational chat requires near-instant validation to maintain conversational flow. However, multimodal processing is computationally expensive. A typical pipeline involving image captioning (e.g., via BLIP-2 [12]) and audio transcription (via Whisper [14]) followed by LLM reasoning introduces significant latency. Current systems struggle to complete this full fusion pipeline in under 3 seconds without sacrificing model size and, consequently, detection accuracy.

Privacy and Deployment Constraints

The most powerful multimodal models (e.g., GPT-4V or Gemini) are primarily cloud-hosted. For highly regulated industries like healthcare, finance, or defence, sending internal chat logs and screenshots to a public cloud API is often unacceptable due to data leakage risks. There is a significant research gap in the development of high-performance **Small Language Models (SLMs)** capable of running effectively on-premise.

Future Directions

Video Stream Analysis

While current research effectively handles static images and short audio clips, real-time video conferencing (e.g., Zoom or Microsoft Teams) remains largely unmoderated regarding content policy. Future architectures must integrate intelligent frame-sampling techniques to moderate live video streams for sensitive visual data without overwhelming network bandwidth[8].

Federated Learning for Privacy

To resolve the conflict between privacy and model improvement, Federated Learning offers a promising path. Future work should explore architectures where organizations can train moderation adapters on their local, sensitive data without sharing raw chat logs with a central model provider. This would enable the system to learn organization-specific slang, norms, and context securely [15].

VI. CONCLUSION

As digital communication increasingly extends beyond text to include images, videos, audio, and mixed-media content, enterprise content moderation systems must evolve to address the growing complexity of multimodal information. This survey examined the current landscape of multimodal content moderation architectures, highlighting the techniques, frameworks, and deployment strategies used to detect harmful, unsafe, or policy-violating content across diverse data modalities.

The analysis demonstrates that multimodal approaches consistently outperform single-modality systems by leveraging complementary signals from text, visual, and audio sources. Advances in deep learning, transformer-based architectures, multimodal fusion techniques, and large foundation models have significantly improved moderation accuracy and contextual understanding. However, challenges remain in areas such as scalability,

explainability, bias mitigation, privacy preservation, cross-cultural interpretation, adversarial robustness, and real-time processing requirements within enterprise environments.

REFERENCES

1. H. Inan, K. Upasani, J. Chi, *et al.*, "Llama Guard: LLM-based input output safeguard for human-AI conversations," *Meta AI Research*, 2023.
2. S. Chennabasappa, C. Nikolaidis, D. Song, *et al.*, "LlamaFirewall: An open-source guardrail for secure AI agents," *arXiv:2501.xxxxx*, 2025.
3. K. Palla, J. L. R. Garc´ıa, *et al.*, "Policy-as-prompt: Rethinking content moderation in the age of large language models," in *Proc. ACM FAccT*, 2025.
4. N. Rodis, C. Sardianos, G. T. Papadopoulos, *et al.*, "Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions," *IEEE Access*, vol. 12, pp. 45678– 45702, 2024.
5. Y. Jin, M. Choi, G. Verma, J. Wang, and S. Kumar, "MM-SOC: Benchmarking multimodal large language models in social media platforms," in *Findings of ACL*, 2024.
6. J. Zheng, X. Ji, and Y. Lu, "RSafe: Incentivizing proactive reasoning to build robust and adaptive LLM safeguards," *arXiv:2501.xxxxx*, 2025.
7. Y. Zhang, M. Li, W. Han, and Y. Yao, "Safety is not only about refusal: Reasoning-enhanced fine-tuning for interpretable LLM safety," *arXiv:2501.xxxxx*, 2025.
8. G. A. Koushik, D. Kanojia, and H. Treharne, "Towards a robust framework for multimodal hate detection: A study on video vs. image-based content," in *Proc. ACM WWW Companion*, 2025.
9. D. Kumar, Y. AbuHashem, and Z. Durumeric, "Watch your language: Investigating content moderation with large language models," *arXiv:2408.xxxxx*, 2024.
10. M. Nasser, N. I. Arshad, I. Nafea, *et al.*, "A systematic review of multimodal fake news detection on social media using deep learning models," *Results in Engineering*, vol. 26, 2025.
11. Y. Dong, R. Mu, X. Huang, *et al.*, "Building guardrails for large language models," in *Proc. ICML*, vol. 235, 2024.
12. C. Gehweiler and O. Lobachev, "Classification of intent in moderating online discussions: An empirical evaluation," *Decision Analytics Journal*, 2024.
13. L. Kearns, "Content moderation assistance through image caption generation," *Intelligent Systems with Applications*, 2024.
14. E. L. T. Tchokote and E. F. Tagne, "Effective multimodal hate speech detection on Facebook hate memes dataset using incremental PCA, SMOTE, and adversarial learning," *Machine Learning with Applications*, vol. 18, 2024.
15. I. T. Ahmed, B. T. Hammad, and N. Jamil, "Image steganalysis based on pre-trained convolutional neural networks," in *Proc. IEEE CSPA*, 2022.