Responsible AI-Driven Adaptive Content Moderation for OTT Platforms: Enabling Inclusive and Family-Safe Streaming in India

Author: Abdeali Makda
Track: Ethics, Equity, and Responsible AI in Sustainability
Contact: 2025.abdealim@isu.ac.in

Abstract

The rapid expansion of Over-The-Top (OTT) streaming platforms has transformed media consumption patterns, particularly in India where television and digital streaming remain central to family-oriented viewing. Yet the increasing prevalence of unfiltered adult content—explicit language, nudity, sexual scenes, substance use, and graphic violence—has made many households reluctant to watch otherwise high-quality films and series together, raising concerns about inclusivity, child safety, and cultural sustainability.

This paper proposes a Responsible AI-driven adaptive content moderation framework that enables age-appropriate, user-controlled personalization of OTT content without compromising the core narrative or creative vision of filmmakers. The system integrates computer vision, natural language processing (NLP), and context-aware Large Language Models (LLMs) to detect, classify, and selectively adapt sensitive content in near real time, while an AI-generated context-preservation layer maintains narrative continuity through brief textual or audio summaries.

Rather than enforcing platform-level censorship, the architecture centers viewer agency, transparency, and cultural sensitivity. Empirical literature, industry data, and existing AI moderation technologies are synthesized to demonstrate how such a system can broaden audience reach, restore family co-viewing practices, strengthen user trust, and contribute to the "Social" pillar of ESG by promoting digital well-being and Responsible AI.

The paper concludes with a roadmap for implementing ethically sustainable content adaptation in diverse cultural contexts.

Introduction

For many Indian families, watching television together in the living room has long been a daily ritual: dinner plates on the center table, grandparents in their designated chairs, children sprawled on the floor, and everyone negotiating what to watch. As affordable smartphones, smart TVs, and broadband spread across India, this ritual migrated to OTT platforms like Netflix, Amazon Prime Video, Disney+ Hotstar, SonyLIV, JioCinema and others. India’s OTT audience is now estimated at around 547–600 million viewers, representing roughly 38–41% of the population and growing steadily.

However, the new “digital living room” is increasingly uncomfortable for families. High-quality series and films frequently include explicit sex scenes, partial or full nudity, graphic violence, abusive language, or substance use as part of realistic storytelling or to appeal to young adult audiences. Parents often find themselves lunging for the remote, scrambling to mute or fast-forward at the last second, or abandoning a show altogether. A content analysis of Indian OTT web series notes that profanity, sexual content, depiction of alcohol and drugs, and violence are now pervasive in mainstream streaming originals, making “watching with family members harder these days” due to vulgarity and offensive language.

The tension is not unique to India. Globally, actors and creators openly acknowledge that their own children cannot watch much of their work. Bob Odenkirk has said he would not allow his kids to watch “Breaking Bad” when they were young, remarking that “you can’t have little kids around while watching Breaking Bad, or the authorities might get involved”. Many performers in comedy and drama, including those working on edgy streaming content, report similar dilemmas: successful shows that define their careers are not suitable for family viewing, and certainly not for younger children. This mirrors what many Indian parents experience when they try to watch critically acclaimed but explicit Indian or international web series with their families.

At the same time, Indian OTT consumption is intense and heavily youth-driven. Studies indicate that Indian teens spend on average 8 hours 29 minutes per day on OTT platforms—far above the global average of 6 hours 45 minutes. OTT addiction prevalence among Indian youth populations has been reported as high as 68% in some cohorts, with binge-watching late into the night associated with physical, mental, and social health issues. Families thus confront a dual challenge: on the one hand, excessive, unsupervised solo OTT viewing by children and adolescents; on the other, the erosion of safe, shared family viewing spaces due to explicit, uncensored content.

The industry-level picture magnifies these concerns. India’s OTT market is one of the fastest-growing in the world, with revenue expected to reach several billion dollars within this decade, and with over 50 platforms competing for attention. India now has over 500M+ OTT users with penetration forecast to cross 600M users by 2029. A significant share of OTT content is R-rated or adult-focused: one analysis of Netflix's Top 10 charts across five countries found that R-rated content accounts for an average of 54% of the catalog, with Indian platforms like ALTBalaji explicitly programming almost all originals for 18+ audiences. Parallelly, a range of "adult-only" OTT apps such as Ullu, Fliz Movies, HotShots, and Kooku have proliferated, monetizing "bold" and sexually explicit regional content.

Unsurprisingly, public anxiety over adult content on OTT platforms is high. A LocalCircles survey reported by Indian news media found that a majority of respondents were concerned about adult content on OTT services such as Netflix, Hotstar, and Amazon Prime and supported stronger censorship or regulatory oversight. Governments have begun reacting: India has issued orders to block 22–25 OTT platforms (mostly fringe "obscene" content providers like Ullu, Rabbit, and others) for streaming sexually explicit content, signaling that decency standards and public morality will be more strictly enforced in the digital sphere. Yet blanket bans or top-down censorship are blunt instruments that risk stifling creative freedom, limiting adult autonomy, and pushing audiences towards piracy or unsafe spaces.

This paper is grounded in a different premise: the problem is not that explicit or mature storytelling exists, but that families lack fine-grained, user-centric tools to adapt such content to their own values, ages, and contexts. Instead of asking "Should this show exist?" The more constructive question is "Can this show be safely and meaningfully watched by different audiences through responsible, personalized adaptation?"

Advances in artificial intelligence—especially computer vision, speech and audio processing, and context-aware Large Language Models (LLMs)—make it technically feasible to identify, label, and adapt sensitive segments of video in near real time. Existing AI content moderation systems already detect nudity, violence, hate symbols, and profanity with high accuracy; for instance, commercial APIs like Amazon Rekognition and others report around 80–90% accuracy in detecting explicit material. Platforms such as YouTube and Facebook have successfully deployed ML-based moderation systems that automatically flag and remove violent extremist content, with up to 98% of such videos being detected algorithmically. Specialized AI vendors such as Hive and WebPurify provide low-latency NSFW moderation and video screening for streaming platforms, capable of blurring or blocking explicit content as it appears in live or on-demand streams.

However, current systems are typically binary—either block/remove the content or allow it—and are optimised for platform-level safety and regulatory compliance, not for nuanced, viewer-controlled adaptation. They also rarely include mechanisms to preserve narrative continuity when segments are skipped or redacted.

This paper proposes an AI-driven adaptive content moderation framework tailored for OTT streaming that:

- Detects sensitive visual, audio, and textual content (nudity, sexual scenes, violence, profanity, substance use) using a multi-stage AI pipeline;
- Applies customizable, user-selected interventions (blurring, muting, bleeping, optional scene skipping) rather than enforcing a single censored version;
- Uses LLM-driven context-preservation layers to generate brief summaries or transitions where segments are skipped, preventing narrative confusion;
- Embeds Responsible AI principles—fairness, transparency, accountability, privacy, and sustainability—into the system’s design, governance, and evaluation, aligning with the conference track “Ethics, Equity, and Responsible AI in Sustainability”.

By placing control in the hands of users, rather than regulators or platforms alone, the framework seeks to restore the possibility of shared, family-friendly viewing without undermining artistic integrity. It also offers a path for OTT platforms to expand their addressable audience, reduce churn from family segments, and demonstrate leadership on ESG-aligned digital well-being.

Review of Literature

OTT Adoption and Consumption Patterns

The global OTT ecosystem has undergone rapid expansion, disrupting traditional television and cinema. Thematic reviews of OTT adoption highlight motivations such as convenience, content variety, personalization, binge-watching affordances, and social interaction. In India, the rise of affordable smartphones, low-cost data, and COVID-19 lockdowns significantly accelerated this shift.

Key trends relevant to this research include:

- High penetration and user base: India’s OTT audience is estimated between 547.3 million and 601.2 million people, representing around 38–41% of the population. Reports project further growth to 630M+ users by 2029.
- Time spent and binge-watching: Indian teens lead the world in OTT usage duration, with an average of 8h 29m per day, far above the global average of 6h 45m. Binge-watching entire seasons in one sitting is common, especially among young adults.

- Device and context: While OTT consumption is often perceived as a solitary, mobile-first habit, the rise of connected TVs and smart TVs is bringing streaming back into the shared living room. India's Connected TV audience is estimated at 129.2 million users, with 35–40 million connected TV homes and around 45 million connected TV sets, making India the third-largest CTV market globally. This has direct implications for family co-viewing and exposure to explicit content on large screens.

- Demographics: OTT usage cuts across age groups, with Generation X and older cohorts increasingly adopting streaming services during and after the pandemic. However, youth and young adults remain the most intensive users.

These patterns suggest both opportunity and risk: OTT is now deeply embedded in everyday life and across generations, but mechanisms for safe, inclusive shared viewing have not kept pace.

Explicit Content on OTT Platforms

Several studies and industry reports highlight the increasing prevalence of explicit themes on OTT platforms:

- A content analysis of Indian shows on major OTT platforms (Netflix, Disney+ Hotstar, Prime Video) documented frequent occurrences of sexual content, implied sexual encounters, violence, drug and alcohol use, and extensive profanity in Hindi. The authors note that this trend makes family co-viewing uncomfortable and may negatively influence young viewers' attitudes and behaviour.

- An Inc42 analysis of Netflix's Top 10 charts across five countries found that the average share of R-rated content was 54%, with Indian content skewing heavily towards adult themes.

- Local Indian platforms such as ALTBalaji, Ullu, MX Player, and Addatimes have built much of their growth on "bold, mature content," with ALTBalaji reportedly targeting almost all of its originals at an 18+ audience.

- The proliferation of adult-only streaming apps (Fliz Movies, HotShots, Kooku and others) targeting regional audiences further indicates strong commercial incentives for explicit content.

While explicit content can contribute to narrative realism, sexual liberation, and more honest portrayals of social issues, there are legitimate concerns about the normalization of violence, objectification, and regressive gender norms, especially when content is consumed privately and without discussion.

Public Concerns, Regulation, and Family Viewing

Surveys and commentary indicate that many viewers are uncomfortable with the volume and intensity of adult content on mainstream OTT platforms:

- A LocalCircles poll with over 40,000 responses found that a majority of Indian users were worried about adult content on OTT platforms and supported some level of censorship or regulation; key concerns included explicit sex scenes, abusive language, and lack of effective parental controls.

- Policy analyses note that OTT regulations in India are evolving, with measures such as the 2023 OTT Rules on tobacco depiction and more recent orders blocking platforms for "obscene and pornographic content". This has created regulatory uncertainty, but also pressure on platforms to demonstrate stronger self-regulation and content classification.

- Qualitative accounts highlight the erosion of family co-viewing: viewers report that "watching with family members is becoming harder" because web series frequently use vulgar and offensive language and feature explicit scenes.

However, existing regulatory responses are primarily platform-centric (blocking, stricter content codes) rather than user-centric. They do not differentiate between households with different values, nor do they allow adults who are comfortable with adult content to continue watching unmodified material while enabling safer variants for children or mixed-age groups.

Psychosocial Impacts and Digital Well-Being

Research on OTT and web streaming addiction among Indian youth reveals concerning patterns:

- A study on OTT addiction found that 68.37% of participants met criteria for OTT addiction, with significant associations between addiction and viewing duration, type of content (often thrillers and explicit material), and late-night binge-watching.

- A comprehensive review of the psychosocial and sleep effects of web streaming on Indian youth identified associations between heavy streaming and sleep deprivation, anxiety, depression, social isolation, and academic difficulties. The authors emphasize the need for media literacy, responsible digital consumption, and policy responses that balance access with well-being.

These findings reinforce the importance of designing content ecosystems that not only entertain but also support healthy consumption patterns and protect vulnerable users.

AI-Based Content Moderation Technologies

AI-driven content moderation has matured significantly in recent years, particularly for social media and user-generated content. Key approaches include:

- Image and video moderation: Computer vision models detect nudity, partial nudity, sexual activity, violence, blood, weapons, and other categories. Services such as Amazon Rekognition, Hive, and WebPurify provide APIs to screen images and videos for explicit or unsafe content, with reported accuracies of around 80–90% in detecting explicit material.

- Text and audio moderation: NLP models classify text and transcripts for profanity, hate speech, harassment, self-harm, and other harmful categories, while audio models detect abusive language and slurs in spoken content.

- Hybrid human-AI workflows: Most practical systems combine automated detection with human review for borderline cases or appeals, recognizing that AI struggles with context, cultural nuance, and satire.

- Specialized models: Advanced systems also incorporate OCR for embedded text in video frames, demographic attribute detection (age, gender) to enforce age compliance, and deepfake detection to identify manipulated or synthetic media.

These technologies are already deployed at scale: YouTube’s ML algorithms flag the vast majority of violent extremist videos before they are widely viewed, and content ID systems match copyrighted content with high precision. However, the primary use case has been removal or demonetization—not adaptive personalization or narrative-preserving modification.

Responsible AI, Ethics, and Sustainability

Responsible AI has emerged as a guiding framework for ethical and trustworthy AI deployment across sectors. Core principles widely recognized in industry and policy discussions include:

- Fairness and inclusion: AI systems should avoid unjust bias and provide equitable outcomes across demographic groups.

- Transparency and explainability: Users and stakeholders should be able to understand how AI systems make decisions and what their limitations are.

- Accountability and governance: Clear lines of responsibility, auditability, and oversight mechanisms should govern AI development and deployment.

- Privacy and security: AI must protect personal data, respect consent, and comply with data protection norms.

- Reliability, safety, and robustness: AI should perform consistently across scenarios and be resilient to adversarial inputs and failure modes.

- Sustainability: Emerging frameworks explicitly include sustainability, emphasizing the environmental footprint of AI and its role in supporting social and economic sustainability.

For media and content moderation, Responsible AI also entails nuanced trade-offs between free expression, harm reduction, cultural pluralism, and user autonomy. Overly aggressive algorithms risk over-censorship and chilling effects on artistic expression, while lax systems may expose users—especially children and marginalized groups—to harm.
This paper situates the proposed adaptive moderation framework within this Responsible AI landscape, focusing specifically on the intersection of Ethics, Equity, and Responsible AI in Sustainability.

Research Objectives and Questions

Building on the literature and real-world context described above, the study pursues the following objectives:

1. To analyze the scope and nature of explicit content on OTT platforms in India and its implications for family co-viewing and digital well-being.

2. To examine existing AI-based content moderation technologies and identify gaps in current approaches with respect to personalization, narrative continuity, and user agency.

3. To conceptualize a Responsible AI-driven, adaptive content moderation framework tailored to OTT streaming, with emphasis on age-appropriate personalization rather than platform-level censorship.

4. To articulate how such a framework can align with and advance the themes of Ethics, Equity, and Responsible AI in Sustainability, particularly in the ESG "Social" pillar.

Corresponding research questions include:

- RQ1: What is the prevalence and character of adult/explicit content on OTT platforms, and how does it affect family viewing practices in India?

- RQ2: What are the capabilities and limitations of current AI models and APIs for detecting and moderating explicit audiovisual content?

- RQ3: How can a multi-stage AI pipeline be designed to enable user-controlled adaptation (blur/mute/skip) while preserving narrative coherence?

- RQ4: What Responsible AI principles and governance mechanisms are needed to ensure that adaptive moderation is ethical, inclusive, and socially sustainable?

Methodology

Given the conceptual and design-oriented nature of the proposed framework, the study adopts a mixed qualitative–conceptual methodology comprising:

1. Systematic literature review: Academic and industry literature on OTT adoption, explicit content prevalence, psychosocial impacts, AI content moderation, and Responsible AI frameworks were reviewed using keyword-based searches (e.g., "OTT India content analysis," "violence and nudity streaming India," "AI video moderation nudity profanity," "Responsible AI framework fairness transparency") across databases and specialized reports.

2. Policy and media analysis: Regulatory developments (OTT rules, platform bans), surveys (LocalCircles), and industry reports (PwC, Deloitte, Ormax Media, MPA, Inc42, etc.) were analyzed to understand the socio-regulatory context and public concerns.

3. Technology landscape scanning: Documentation and whitepapers from commercial AI moderation providers (Amazon Rekognition, Hive, WebPurify, and others) and technical reviews on AI moderation were examined to synthesize capabilities, accuracy claims, and deployment patterns.

4. Conceptual system design: Drawing on the above insights, a multi-stage AI architecture for adaptive OTT moderation was designed, including detection, classification, adaptation, and context-preservation layers, and mapped to Responsible AI principles.

The study does not collect primary empirical data (e.g., user experiments) but proposes a framework and outlines how future empirical studies could evaluate its effectiveness.

Data and Contextual Analysis

OTT Penetration and Usage in India

Metric	Value (approx.)
Total OTT viewers in India (2024)	547.3 million (≈38% population)
Broader OTT audience universe (2025 est.)	601.2 million (41.1% population)
Projected OTT users by 2029	634 million (penetration ~42%)
Active OTT paid subscriptions (India)	148 million
Average daily time spent by Indian teens on OTT	8h 29m (vs 6h 45m global avg)
Smart/Connected TV OTT audience	129.2 million users; 35–40M homes

These figures underscore three points:

- OTT has moved from niche to mainstream, with nearly two in five Indians using streaming services.
- Youth and teens are heavy users, amplifying concerns about unmoderated exposure.
- The living room is re-emerging as a key OTT consumption context, making family-safe viewing more salient.

Public Perception and Demand for Moderation

A survey by LocalCircles reported in Indian media indicates that:

- A majority of respondents are concerned about adult content (sex, nudity, profanity) on OTT platforms;

- Many support regulatory steps or censorship to prevent children from access and to uphold cultural norms.

Parallelly, global and Indian discourse around digital well-being stresses the need for tools that allow parents to manage what children watch and how long they watch, without necessarily banning platforms outright.

Limitations of Current Platform Tools

Major OTT platforms currently offer basic content controls:

- Age-rating labels and maturity ratings;

- Profile-level restrictions (kids vs general);

- Parent PINs and viewing restrictions by rating category.

However, these controls are coarse-grained. They do not:

- Allow selective filtering of specific categories (e.g., allow mild violence but block nudity);

- Provide per-household or per-profile customization beyond broad age ratings;

- Offer dynamic adaptation within a show (e.g., blur a few seconds of nudity while retaining the rest of the scene);

- Support narrative-preserving skip mechanisms with AI-generated summaries.

As a result, families must either avoid certain shows entirely, risk exposure to unwanted content, or perform manual “remote-control censorship” that is stressful and error-prone.

Technical Landscape: Existing AI Models and Systems

Computer Vision for Visual Content Moderation

Modern computer vision models, often based on convolutional neural networks (CNNs) and transformer architectures, can detect:

- Full and partial nudity, including specific body regions;
- Sexual acts and suggestive poses;
- Violence, blood, and gore;
- Weapons and dangerous objects;
- Hate symbols and extremist imagery.

Commercial APIs such as Amazon Rekognition, Hive, and WebPurify claim detection accuracies of 80–90% for explicit nudity and other categories, and support video frame analysis, sometimes with low latency suitable for streaming. These APIs typically output:

- Category labels (e.g., “Explicit Nudity,” “Violence,” “Suggestive,” “Safe”);
- Confidence scores;
- Bounding boxes for sensitive regions within frames.

NLP and Audio Models for Profanity and Dialogue

Text and audio moderation uses NLP and speech recognition to:

- Transcribe dialogue into text (via ASR—automatic speech recognition);
- Classify segments of text for profanity, slurs, hate speech, self-harm content, or sensitive topics;
- Optionally detect sentiment, toxicity, or emotional intensity.

Profanity filters can identify swear words and abusive phrases in multiple languages and can be configured to either mute audio, bleep specific words, or replace them with milder alternatives in subtitles.

Hybrid Moderation Architectures

Most large-scale platforms implement hybrid pipelines:

- AI models perform initial detection and triage, flagging likely violations.

- Human moderators review flagged content for context and make final decisions in ambiguous cases.

- Feedback from human decisions is used to retrain and improve models over time.

Proposed Responsible AI-Driven Adaptive Content Moderation Framework

Design Goals:

The proposed framework is built around four primary goals:

1. User agency and personalization: Allow viewers (or parents/guardians) to define what types and intensity of content they are comfortable with, instead of imposing a uniform censorship standard.

2. Narrative integrity: Preserve the core story, themes, and artistic intent, even when specific scenes are blurred, muted, or skipped.

3. Responsible AI and ethics: Ensure transparency, fairness, privacy, and accountability in how AI models detect and adapt content.

4. Scalability and interoperability: Design the system so it can operate across multiple OTT platforms, languages, and genres.

Detection Layer (Computer Vision + Audio + NLP):

- Apply visual models to detect nudity, sexual acts, imagery, etc.

- Use ASR to transcribe audio; apply text moderation models to detect profanity, slurs, or sensitive topics in dialogue.

- Assign labels to micro-scenes (e.g., “sexual content, high intensity,” “moderate violence,” “strong language”).

Classification and Policy Layer:

- Map detected labels to standardized content categories and severity levels (e.g., mild/moderate/severe).

- Compare against user-defined preference profiles that specify which categories and intensity levels are allowed, blurred, muted, or skipped.

- Example: A "Family Mode" profile might permit mild romantic scenes but blur nudity, mute strong profanity, and optionally skip explicit sexual scenes.

Adaptation Layer:

- Visual adaptation: Apply dynamic blurring or masking to sensitive regions within frames (based on bounding boxes), or crop frames to reduce exposure while preserving as much context as possible.

- Audio adaptation: Mute or bleep specific words, reduce volume during intense scenes, or replace dialogue segments in subtitles with neutral alternatives.

- Structural adaptation: For segments that exceed thresholds (e.g., explicit sex scenes), mark them for skipping and adjust playback timelines accordingly.

Context-Preservation Layer (LLM-based):

- When a segment is skipped, use an LLM to generate a concise textual or audio summary that conveys essential plot information, character development, and emotional beats without explicit detail.

- Summaries are displayed as on-screen text, a brief narration, or a short overlay card ("While this scene is skipped as per your settings, the following happens: …").

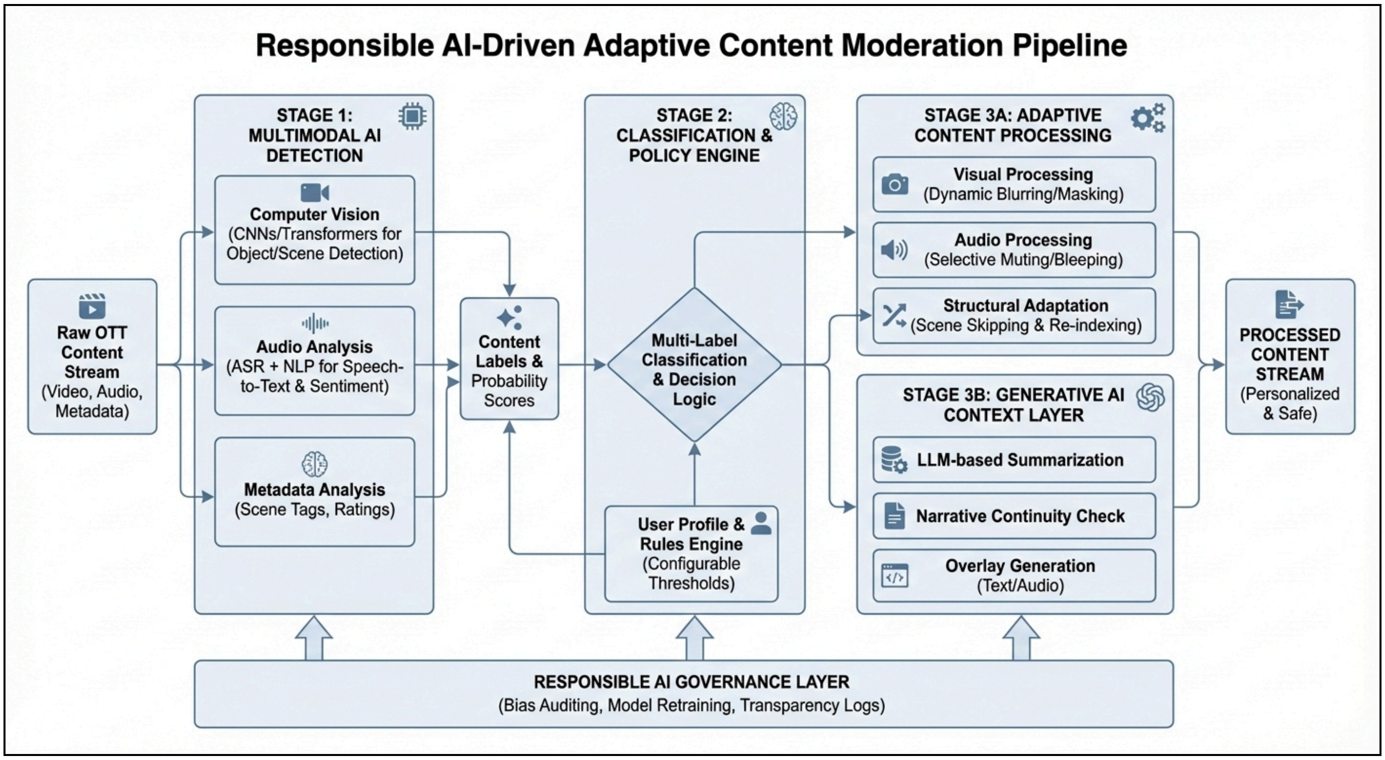
- The LLM uses script, subtitles, and metadata as input to ensure accuracy and continuity.

User Interface and Controls:

- Simple, transparent settings for users to choose modes (e.g., "Unrestricted," "Family," "Kids," "Custom") and to see what each mode does.

- Real-time toggles to temporarily relax or tighten filtering.

- Clear indicators when AI intervention occurs (e.g., small icon when content is blurred or skipped).


Responsible AI-Driven Adaptive Content Moderation Pipeline
Raw OTT Content Stream (Video, Audio, Metadata)
STAGE 1: MULTIMODAL AI DETECTION
Computer Vision (CNNs/Transformers for Object/Scene Detection)
Audio Analysis (ASR + NLP for Speech-to-Text & Sentiment)
Metadata Analysis (Scene Tags, Ratings)
Content Labels & Probability Scores
STAGE 2: CLASSIFICATION & POLICY ENGINE
Multi-Label Classification & Decision Logic
User Profile & Rules Engine (Configurable Thresholds)
STAGE 3A: ADAPTIVE CONTENT PROCESSING
Visual Processing (Dynamic Blurring/Masking)
Audio Processing (Selective Muting/Bleeping)
Structural Adaptation (Scene Skipping & Re-indexing)
STAGE 3B: GENERATIVE AI CONTEXT LAYER
LLM-based Summarization
Narrative Continuity Check
Overlay Generation (Text/Audio)
PROCESSED CONTENT STREAM (Personalized & Safe)
RESPONSIBLE AI GOVERNANCE LAYER (Bias Auditing, Model Retraining, Transparency Logs)

Example User Scenarios

1. Family movie night: Parents enable “Family Mode” with blurring of nudity and muting of strong profanity. When an explicit scene appears in a popular series, the system detects it in real time, automatically skips it, and shows a brief summary: “The characters confess their love and decide to move in together.” The story continues seamlessly, and no one has to scramble for the remote.

2. Teen profile with partial restrictions: A 16-year-old viewer’s profile allows mild violence and romantic scenes but not explicit sex or heavy drug use. The system leaves most action sequences untouched but skips a graphic sexual encounter, replacing it with a textual summary.

3. Adult viewer with cultural preferences: An adult viewer may be comfortable with violence but uncomfortable with profanity or nudity when watching with elders. They configure a custom profile to only mute profanity and blur nudity, but not skip scenes entirely, preserving as much of the original pacing as possible.

Ethics, Equity, and Responsible AI in Sustainability

A core contribution of this paper is to explicitly map the proposed framework onto the conference track themes: Ethics, Equity, and Responsible AI in Sustainability.

Ethics: Balancing Artistic Integrity, Autonomy, and Harm Reduction

The framework avoids one-size-fits-all censorship. Instead, it:

- Respects artistic integrity by leaving the original master content intact; AI adaptation occurs at the playback layer, under user control. Creators can continue to craft mature, complex stories without being forced into homogenized "family-safe" versions.

- Enhances user autonomy by giving households granular control over what content is filtered and how. Adults can choose unfiltered experiences; parents can tailor settings for children; elders can avoid content that clashes with their values.

- Reduces harm by making it easier to protect children and sensitive viewers from sudden exposure to explicit material, aligning with principles of non-maleficence and care.

Ethically, this shifts power from opaque content gatekeepers to informed users, while building in transparent communication about when and why AI intervenes.

Equity: Inclusion Across Cultures, Languages, and Socioeconomic Strata

Equity considerations are central in two ways:

1. Cultural and linguistic fairness:

- Models must perform consistently across Indian languages (Hindi, Tamil, Telugu, Bengali, Marathi, etc.) and diverse cultural norms. Profanity and taboo topics differ by region; training data and human review teams must reflect this diversity to avoid bias.

- The system should support regional OTT content, not just English/Hindi mainstream shows, to ensure that rural and non-metro audiences benefit equally from adaptive moderation.

2. Access and affordability:

- Lightweight model deployment and edge optimization (e.g., on-device pre-processing or using CDNs) can reduce bandwidth and compute costs, making the feature available on low-end devices and in lower-income households.

- Platforms may choose to offer adaptive moderation at no extra cost as part of their ESG commitments, avoiding a situation where only premium tiers can access safer, family-friendly features.

By designing for diverse users, the framework supports equitable access to high-quality content without forcing marginalized groups to choose between cultural relevance and safety.

Responsible AI: Principles and Governance

The system is explicitly aligned with established Responsible AI principles:

1. Fairness:

- Regular audits for differential error rates across genders, age groups, dialects, and content genres.

- Bias mitigation techniques in training data and model tuning, especially for detecting nudity and sexual content in ways that do not disproportionately penalize certain bodies or cultures.

2. Transparency and explainability:

- Clear explanations in the UI of what each setting does and which categories are being filtered.

- Optional “Why was this scene modified?” tooltips summarizing the detected content category.

3. Accountability and governance:

- Documented model cards and data sheets for detection models, disclosing known limitations.

- Internal governance structures within OTT platforms (or third-party moderation providers) to oversee model updates, handle grievances, and ensure compliance with laws.

4. Privacy and security:

- Minimizing collection of personal viewing data; using aggregation and anonymization for model improvement.

- Ensuring that sensitive content flagged by the system is not exposed to unnecessary human review unless required (e.g., for legal reasons or appeals), and that any stored snippets are securely protected.

5. Reliability and safety:

- Continuous monitoring of false positives (over-filtering) and false negatives (missed harmful content), with human-in-the-loop corrections and model retraining.

- Robustness testing against adversarial inputs (e.g., attempts to bypass detection via overlays or filters).

Sustainability: Social and Environmental Dimensions

From an ESG standpoint, the framework primarily strengthens the Social pillar by:

- Promoting digital well-being and healthy media habits, especially for children and youth who are at risk of OTT addiction and overexposure to explicit content.
- Supporting family cohesion and intergenerational bonding by making it possible to watch the same content together, with customization for shared comfort levels.

- Enhancing trust in digital media ecosystems, potentially reducing backlash, bans, or regulatory overreach by demonstrating proactive, user-centric self-regulation.

On the environmental side, AI-based systems do consume computational resources. Responsible design should therefore:

- Use efficient model architectures and inference strategies to minimize energy usage;

- Explore shared models across platforms, or federated learning approaches, to reduce duplication;

- Balance the environmental cost of AI computation against the social benefits of safer, more inclusive content ecosystems.

Overall, the framework offers a model of sustainable AI: mitigating harms while enabling continued innovation and creative expression in the digital media industry.

Discussion and Implications

Implications for Families and Viewers

For families, the proposed system promises:

- Reduced stress and "remote-control anxiety" when watching content with children or elders;

- The ability to discover and enjoy acclaimed series and films that were previously avoided due to a few explicit scenes;

- A more transparent sense of control and safety, fostering trust in OTT platforms.

However, user education will be critical. Viewers must understand that AI filters are not perfect, that some explicit content may still appear, and that no system replaces parental judgment.

Implications for OTT Platforms

- Adaptive moderation can become a differentiated feature, attracting family segments and older audiences who currently stick to traditional TV due to explicit OTT content.

- It can mitigate reputational and regulatory risks by demonstrating Responsible AI practices and proactive content safety measures, possibly influencing regulatory frameworks in a favourable way.

- It opens up opportunities for tiered personalization, including "cultural modes" tuned to specific regions and languages.

Conclusion

OTT platforms have reshaped how India and the world consume media, bringing unprecedented choice, convenience, and creative diversity. Yet the same platforms have made shared, family-friendly viewing more difficult due to the rising prevalence of explicit content, from nudity and sexual scenes to graphic violence and pervasive profanity. Survey data, content analyses, and regulatory responses all point to growing concern among viewers and policymakers about the social consequences of unfiltered streaming.

At the same time, advances in AI content moderation—spanning computer vision, NLP, and LLMs—create a feasible path toward more nuanced solutions. Existing systems already detect explicit content with high accuracy and are deployed at scale in social media and other platforms.

This paper proposes a Responsible AI-driven adaptive content moderation framework for OTT platforms that:

- Uses multi-stage AI detection to identify sensitive visual and audio content;
- Applies user-configurable interventions (blur, mute, skip) instead of blunt censorship;
- Employs LLM-based context-preserving summaries to maintain story comprehension when scenes are skipped;
- Embeds Responsible AI principles of fairness, transparency, accountability, privacy, and sustainability into its design and governance.

By centering user agency, ethical safeguards, and cultural sensitivity, the framework offers a way to reconcile artistic integrity with child safety, family cohesion, and digital well-being. It aligns closely with the track themes of Ethics, Equity, and Responsible AI in Sustainability, contributing to the Social dimension of ESG through more inclusive, trustworthy media ecosystems.

Future work should involve prototyping and piloting such systems with real users on selected OTT platforms, conducting empirical evaluations of user satisfaction, perceived safety, and impacts on viewing behaviour, and refining governance models in collaboration with regulators, creators, and civil society. Ultimately, intelligent content adaptation can help ensure that the next phase of the streaming revolution is not just more immersive and personalized, but also more ethical, equitable, and sustainable.

References

Peer-Reviewed Journals and Conference Papers

1. Basu, S., & Das, S. (2024). The online video ecology for preschoolers in India: Investigating the creative industry practices. Asian Journal of Communication. https://doi.org/10.1080/01296612.2024.2429907
2. Singh, R., & Kumar, A. (2011). CRM index development and validation in Indian hospitals. International Journal of Healthcare Delivery Reform Initiatives. https://doi.org/10.4018/jhdri.2011040101
3. Sharma, P. (2024). Revolution in data market: A study on data consumption in India. International Journal of Current Science Research and Review. http://ijcsrr.org/wp-content/uploads/2024/05/89-2805-2024.pdf
4. Gupta, N. (2019). Comparative study of video service providers and piracy. International Journal of Advanced Research. http://www.journalijar.com/uploads/67_IJAR-29684.pdf
5. Lee, J., & Kim, S. (2024). Baby boomers’ Over-The-Top (OTT) rush—Older customers on new platforms. Cogent Arts & Humanities. https://doi.org/10.1080/23311975.2024.2327131
6. Putri, A. R., & Santoso, S. (2023). Examining new media consumption from the standpoint of OTT streaming services. Indonesian Journal of Business Strategy. https://ijbs.petra.ac.id/index.php/ijbs/article/download/303/112
7. Rao, V. (2023). The impact of OTT platforms on the Indian film industry post the Covid-19 pandemic. International Journal of Frontier in Management Research. https://www.ijfmr.com/papers/2023/5/6131.pdf
8. Chen, L. (2011). Statistical information of the increased demand for VOD with mobile devices in Asia. https://arxiv.org/pdf/1112.2042.pdf
9. Johnson, M. (2023). Adoption and usage of over-the-top entertainment services. In the Handbook of Research on Adoption of New Technologies. https://doi.org/10.4018/978-1-6684-7967-4.ch001
10. Patel, K., & Desai, R. (2023). Streaming towards innovation: Understanding consumer adoption of OTT services through IRT and TAM. Cogent Business & Management. https://doi.org/10.1080/23311975.2023.2283917

Industry Reports and Surveys

11. BARC India & Nielsen. (2025). TV highest reach in India: 217M households FY2024. https://www.medianews4u.com/television-remains-the-highest-reach-medium-in-india-with-tv-viewing-households-estimated-at-217-mil

12. Gupta, A., et al. (2025). Over-the-top (OTT) entertainment viewership addiction. https://pmc.ncbi.nlm.nih.gov/articles/PMC12007762/
13. GrowthX Club. (2024). India will have 500M+ OTT users in 2024. LinkedIn Post. https://www.linkedin.com/posts/growthxclub_india-has-500m-ott-users-in-2024-with-activity-7277969378178936832-vQ9y
14. Economic Times. (2025). India has 601 million OTT users and 148 million active paid subscriptions. The Economic Times. https://economictimes.com/industry/media/entertainment/india-has-601-million-ott-users-and-148-million-active-paid-subscriptions
15. ISME. (2021). Impact of OTT media on the family and the individual. https://www.isme.in/impact-of-ott-media-on-the-family-and-the-individual/
16. CNBC TV18. (2020). OTT platform content: Adult content worries many, majority seek censorship, finds survey. https://www.cnbctv18.com/technology/ott-platform-content-majority-seek-censorship-finds-survey-5358391.htm
17. SSCBS. (2025). Over-the-top (OTT) industry report. https://sscbs.du.ac.ac.in/wp-content/uploads/2025/09/OTT-Industry-Report.pdf
18. VisionIAS. (2025). 25 OTT platforms ban: India’s digital content regulation dynamics. https://visionias.in/blog/current-affairs/25-ott-platforms-ban-indias-digital-content-regulation-dynamics

AI Moderation and Responsible AI Sources

19. Inc42. (2022). Violence & nudity: The dark underbelly of OTT in India. https://inc42.com/datalab/violence-nudity-the-dark-underbelly-of-ott-in-india/
20. Label Your Data. (2023). AI content moderation for online responsibility. https://labelyourdata.com/articles/ai-content-moderation
21. Global Tech Council. (2024). What is Responsible AI? https://www.globaltechcouncil.org/artificial-intelligence/responsible-ai/
22. WebPurify. (2024). Video moderation - AI-based and live custom. https://www.webpurify.com/video-moderation/
23. Pegasystems. (2024). What is responsible AI? A complete guide. https://www.pega.com/responsible-ai
24. Hive AI. (2025). Streaming platforms - Hive AI. https://thehive.ai/solutions/streaming-platforms
25. Blue Prism. (2025). 5 principles for Responsible AI. https://www.blueprism.com/guides/ai/responsible-ai/
26. ATIDIV. (2025). Enhance content moderation with AI. https://atidiv.com/enhance-content-moderation-with-ai/

27. Neudesic. (2024). What is Responsible AI: Understanding its importance. https://www.neudesic.com/blog/understand-responsible-ai/
28. Estha AI. (2025). 12 best AI content moderation APIs compared. https://estha.ai/blog/12-best-ai-content-moderation-apis-compared-the-complete-guide/
29. Tredence. (2025). Build trust with Responsible AI frameworks. https://www.tredence.com/blog/responsible-ai-frameworks
30. BuzzFeed. (2022). 10 celebs who don't let their kids watch their work. https://www.buzzfeed.com/abhaahad/celebs-who-wont-let-kids-watch-work
31. Quantum CS. (2021). India watches more frank sexual content, is that a good thing? https://quantumcs.com/sg/india-is-watching-more-frank-sexual-content-but-is-that-a-good-thing/
32. IJCRT. (2024). Content analysis of Indian shows on OTT platforms. https://www.ijcrt.org/papers/IJCRT24A4671.pdf
33. Storyboard18. (2025). Obscenity ban sparks OTT industry shake-up. https://www.storyboard18.com/how-it-works/obscenity-ban-sparks-ott-industry-shake-up-netflix-prime-video-and-others-may-tighten
34. Cureus. (2024). Tobacco imagery in movies and web series streaming in India. https://assets.cureus.com/uploads/original_article/pdf/231898/20240307-16243-biltz8.pdf