Vol.7 Iss.1 2026 Communication & Education Review
https://doi.org/10.37420/j.cer.2026.030

A Review of Automated Content Moderation on Live-Streaming and Short-Video Platforms: From Offline Batch Processing to Real-Time Streaming Detection

Lei Weibin*
Department of Broadcasting, Henan Radio and Television Station
No. 2 Weiyi Road, Jinshui District, Zhengzhou City, Henan Province, China
***Corresponding author, E-mail: lucyboy2022@163.com**

Abstract:

The rapid growth of live-streaming and short-video platforms has led to the large-scale aggregation of user-generated content under conditions of high concurrency, strong interaction, and short dissemination chains, thereby placing content governance under multiple constraints involving detection accuracy, processing latency, and computational cost. This article systematically reviews the evolution of automated content moderation technologies from offline batch processing to real-time streaming detection from three perspectives: processing sequence, modality coverage, and model paradigm. The review shows that early offline moderation approaches based on keywords, rule repositories, and traditional machine learning offered advantages in deployment cost and workflow standardization, but exhibited clear limitations in handling adversarial textual variants, complex semantic understanding, and cross-modal violation recognition. Deep learning methods promoted the transition from rule matching to representation learning and significantly improved unimodal recognition performance for text, images, and speech. Multimodal fusion and vision-language large models further enhanced the modeling of implicit semantics, scene relationships, and policy boundaries, but also introduced greater inference latency, higher resource consumption, and limited interpretability. Based on existing studies and platform practice, a relatively robust engineering path is a cascaded moderation architecture composed of lightweight online filtering, multimodal re-ranking, and human review as a safeguard. Future research should continue to advance lightweight inference for streaming scenarios, cross-domain generalization, explainable decision-making, and the coordination between technical systems and governance mechanisms.

Keywords:

automated content moderation; live-streaming platforms; short-video platforms; multimodal fusion; real-time streaming detection; vision-language large models; human-machine collaboration

1 引言

直播与短视频平台已经成为网络内容生产、分发与消费的重要基础设施。与传统图文平台相比，此类平台上的内容具有持续流入、多模态耦合和强互动反馈等特征，视频画面、语音、字幕、弹幕、评论与用户行为往往在同一时间窗口内共同构成风险语境 [7, 11, 14-15]。在这一背景下，平台审核

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

任务已由发布后的静态检查逐步转向传播过程中的动态识别与前置拦截。

从技术演进看，平台内容审核大体经历了三个阶段。第一阶段以人工审核、关键词过滤与规则引擎为主，主要依赖黑名单、人工经验和模板化匹配。第二阶段以监督学习驱动的文本、图像与语音分类器为代表，通过数据标注和特征学习提升自动化程度。第三阶段进一步走向多模态融合、流式处理与大模型增强，审核系统开始具备一定的跨模态理解、零样本迁移和规则跟随能力 [1-4, 7-11, 15-18]。

尽管技术能力不断提升，内容治理的复杂性并未随之显著降低。其一，违规内容持续演化，同音替换、符号插入、图文错配、语音规避与跨模态拼接等策略不断削弱静态规则与单模态模型的有效性 [1, 6, 13, 19]。其二，直播场景要求系统在毫秒级至秒级时间窗口内完成识别、决策与处置，而多模态重模型与视觉语言大模型通常伴随更高的推理时延和算力开销 [8, 10-11, 16-17, 27]。其三，内容审核并非单纯的分类问题，还涉及规则解释、边界划分、申诉复核与治理责任分配，因而天然具有社会技术系统属性 [6, 12, 20-24, 30]。

基于此，本文围绕直播与短视频平台内容自动审核技术的发展脉络，重点讨论三个问题：离线批处理方案何以演进为面向直播场景的实时流式检测架构；多模态与大模型方法为审核系统带来了哪些能力增益及工程代价；面向未来，自动审核系统在算法、系统与治理层面仍面临哪些关键瓶颈。

2 内容自动审核技术发展与分类

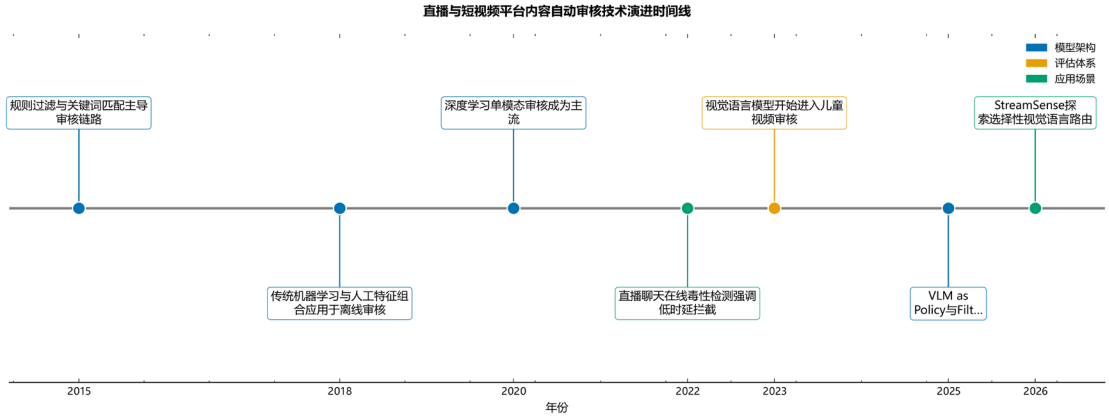


Figure 1. 直播与短视频平台内容自动审核技术演进时间线。

内容自动审核技术可从处理时序、模态覆盖与模型范式三个维度加以分析。按处理时序，可区分为离线批处理与实时流式检测；按模态覆盖，可区分为单模态审核与多模态联合审核；按模型范式，则可概括为规则与传统机器学习、深度学习，以及以大语言模型和视觉语言大模型为代表的大模型驱动方法。上述维度在实际生产环境中通常交叉组合，共同构成平台审核链路。

2.1 离线批处理阶段：规则方法与传统机器学习

离线批处理是平台内容审核最早的大规模工程形态。其典型流程是：内容上传后进入异步队列，再由规则引擎、特征提取模块与分类器进行批量处理，最终将高风险样本送入人工复核环节。该模式适用于点播视频、媒资入库和历史内容回查等对时延要求相对宽松的场景 [2, 4, 13]。

在这一阶段，文本审核主要依赖关键词过滤、词袋模型、TF-IDF 特征及 SVM、随机森林等传统分类器；图像与视频审核则更多采用颜色直方图、局部纹理、显式目标检测等人工特征。其优点在

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

于流程清晰、部署门槛较低，且便于与既有规则库联动；不足在于语义表达能力有限，难以处理文本变体、上下文依赖和组合型违规内容 [1,4,13]。

相关研究表明，在短文本审核任务中，不法内容常通过谐音替换、字符拆分、拼音缩写等方式规避规则系统，导致传统方法在对抗样本上的性能明显下降 [1]。针对中文电商直播内容违规的离线检测研究也表明，即便在相对封闭的业务场景中，单纯依赖离线识别链路也难以及时覆盖语音、画面和互动语境共同构成的复杂风险 [13]。这说明离线批处理虽然奠定了审核流程基础，但其核心能力仍主要停留在局部特征匹配和浅层语义判断层面。

2.2 深度学习与多模态融合阶段

随着算力提升和标注数据积累，内容审核系统逐步由人工特征转向深度表征学习。文本审核开始广泛采用 CNN、RNN、BERT 等模型，图像与视频审核则更多引入 ResNet、时序卷积网络和 Transformer 架构。与传统方法相比，深度学习模型能够学习更稳定的语义特征，在复杂表达识别、上下文建模与视觉场景理解方面表现更优 [1,3-5]。

在真实平台场景中，单模态模型往往难以覆盖复杂违规语境。例如，视频画面本身可能并不构成风险，但在字幕、语音或评论的共同作用下才具有明显违规倾向；又如，一段文本在普通语境下无害，但在特定图像、表演或直播互动情境中可能具有诱导、歧视或擦边属性。因此，多模态联合审核逐渐成为平台治理的重要方向 [2-5,13-15,18]。

现有研究与工程实践表明，多模态系统至少具有三方面优势。其一，它能降低单一模态造成的误判。例如，OCR 识别出的字幕与 ASR 转写的语音内容，可以弥补视觉模型对隐性语义识别的不足。其二，它有助于识别“图文不一致”“语音规避”“场景诱导”等复杂违规模式。其三，多源证据融合可以为人工复核与申诉处理提供更完整的判定依据 [2-5,14,18,29]。不过，多模态系统也面临模态对齐复杂、训练成本高、推理链路长和部署维护困难等问题，尤其在低时延直播场景中，这些问题更为突出 [8,16-18,27,29]。

2.3 实时流式检测阶段

直播场景的兴起改变了内容审核的时间逻辑。与离线批处理不同，流式审核无法等待完整视频生成后再作统一判断，而必须在内容传播过程中持续感知、持续评分并及时处置。由此，审核系统逐步由批处理流水线演化为在线流处理系统 [9-11,16-17,27,29]。

从系统设计看，实时流式检测通常采用分层级联架构。第一层为帧级或句级快速过滤器，使用轻量模型在极短时间内完成粗筛；第二层为片段级精细分类器，对高风险候选样本进行多模态复判；第三层为状态追踪与决策模块，用于整合跨帧、跨句和跨轮互动上下文，从而降低单帧噪声带来的误报 [8,10-11,16-18,27]。已有研究指出，在直播聊天场景中，审核模型不仅要保证检测性能，更要在用户实际读取消息之前完成拦截，这一要求直接决定了系统必须围绕低时延目标进行设计 [10]。

近年来，一些面向流式视频分析和流媒体系统优化的研究，也为实时审核提供了方法启发。在线视频编辑中的时序一致性维护机制，可为流式审核中的状态传递与上下文保持提供参考 [9]；特征预测与帧间复用方法则表明，通过前向估计可以提升流式场景中的信息利用效率 [11]。进一步地，StreamSense 通过选择性视觉语言模型路由减少不必要的重模型调用，MoQ 相关研究则从分布式过滤角度提出了低成本部署路径 [16-17]。这些工作表明，实时审核的关键并不只是提升离线模型的推理速度，而是围绕数据流、时序信息与资源预算重构系统。

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

2.4 大模型驱动的新范式

近两年，大语言模型与视觉语言大模型为内容审核带来了新的方法可能性。一方面，大语言模型在变体文本理解、规则跟随、复杂语义归纳和跨案例知识迁移方面表现出较强能力，可用于处理传统分类器难以覆盖的隐喻表达、擦边表述与上下文依赖场景[1,19-25]。另一方面，视觉语言大模型通过联合建模图像、文本乃至视频片段，增强了审核系统的跨模态语义对齐与规则解释能力[7-8,14-18,27]。

VLM as Policy 是这一方向的代表性工作。该方法不再将审核任务简单视为固定标签分类，而是尝试把平台规则编码为自然语言策略，由视觉语言模型在具体个案中执行解释和判断[15]。这一思路的价值在于，当规则边界调整或新型风险出现时，系统不必完全依赖重新标注与重新训练，即可通过策略更新提升适应速度。与此相近，基于检索增强、跨会话证据聚合和动态规则更新的方法，也开始进入直播风险评估与平台审核研究视野[18-19,27]。

不过，大模型驱动范式并非没有代价。相关研究指出，视觉语言模型虽然具备一定零样本潜力，但在儿童视频等细粒度审核任务上的稳定性仍然有限[7]。Filter-And-Refine 等级联系统则表明，在工业场景中，更现实的方案通常不是让大模型覆盖全部内容，而是通过“先过滤、后精判”的方式将其部署在高风险样本上，以兼顾吞吐量与审核质量[8]。此外，AI 与人工审核的比较研究也显示，大模型在规则一致性和多模态理解方面具有潜力，但在边界个案、价值判断与高风险责任场景中，仍难完全替代人工裁量[20-24,30]。因此，大模型更适合作为审核链路中的高阶推理模块，而非唯一模块。

3 对比分析与关键挑战

3.1 离线批处理与实时流式检测的范式差异

表1 主要内容审核技术范式对比

技术范式	代表路径	模态覆盖	时延水平	主要优势	主要局限
离线批处理	规则+SVM/随机森林	文本/图像单模态为主	分钟级至小时级	流程稳定、部署成本低	泛化能力弱，难适应直播
单模态深度学习	CNN/BERT/ResNet	单模态	百毫秒至秒级	语义表征能力提升	跨模态语境覆盖不足
多模态级联	OCR+ASR+视觉联合	图像+文本+语音	秒级	复杂场景覆盖能力强	链路长、推理成本高
视觉语言大模型	VLM as Policy	视觉+语言	秒级	规则解释与泛化能力强	部署成本高、实时性受限
流式轻量路由	StreamSense/在线过滤	流式多模态	毫秒级至亚秒级	适配直播场景、吞吐高	深层语义理解上限受限

Table 1. 主要内容审核技术范式对比。

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

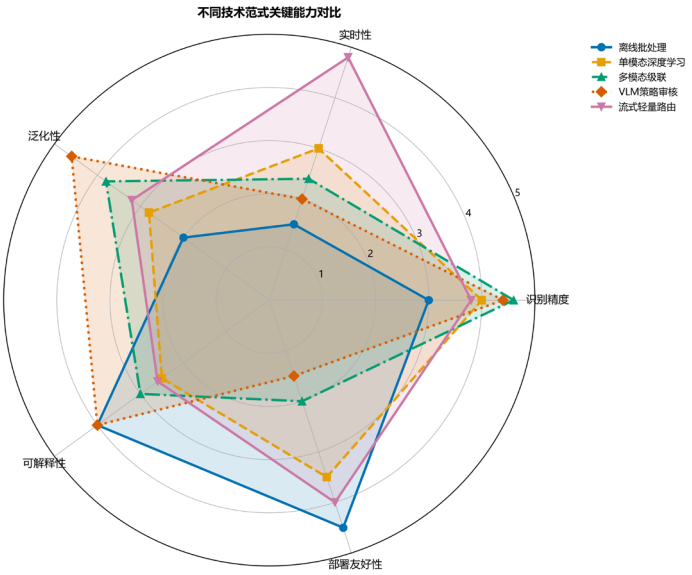


Figure 2. 不同技术范式关键能力对比。

离线批处理与实时流式检测的根本差异，在于系统优化目标不同。前者强调完整性、可回溯性与精细分析，允许对整段视频、全量字幕和多轮上下文进行综合建模，因此更适合部署复杂模型和深层多模态链路；后者强调低时延、连续响应和在线处置，更依赖轻量模型、缓存机制、关键帧采样和分级决策 [8-11, 16-18, 27]。

在检测性能方面，离线方案通常更容易获得较高上限，因为它可以使用更长时窗与更丰富语境；但在实际治理效果上，直播场景中的及时可用往往比事后更准更重要。对于违规直播、实时辱骂、诱导互动和突发性风险内容，一旦错过处置窗口，即使后续识别结果更准确，也难以弥补传播后果 [10, 13, 15, 19, 27]。

在模态使用方面，离线方案更适合执行长视频总结、跨片段语义一致性分析和复杂证据拼接；实时方案则更适合采用“轻量粗筛—疑似样本升级复判”的策略，以避免所有内容均进入重模型链路 [8, 16-18, 27, 29]。因此，两类范式并非简单替代关系，而是在平台治理体系中承担不同职能：前者偏向存量治理与深度复核，后者偏向增量治理与前置拦截。

3.2 当前系统面临的核心挑战

第一，对抗性规避持续升级。无论是文本中的谐音、缩写、错拼与符号插入，还是视频中的遮挡、镜像、语音混淆和图文错配，违规内容都在主动适应平台检测策略，导致静态规则与单一模型不断失效 [1, 6, 13, 18-19]。

第二，实时性与精度之间存在结构性张力。直播审核天然要求低时延，但复杂违规内容的识别通常依赖更长上下文、更深跨模态推理以及更复杂的决策链路。大模型虽具较强语义能力，却也带来更高的显存占用、推理时延与部署成本 [7-8, 10-11, 16-17, 20, 27]。

第三，跨域泛化能力仍然不足。平台风险内容会随热点事件、亚文化表达、地域语言和社会议题迅速演变，导致训练数据与线上分布持续漂移。模型在训练域内表现良好，并不意味着其能够稳定迁移到新场景、新类目与新表达方式 [15, 18-25, 28]。

第四，可解释性与治理协同仍显薄弱。许多审核决策具有边界模糊、价值判断强和申诉成本高等特征，单纯依赖模型分数难以支撑高风险处置。相关研究已指出，自动审核系统必须嵌入规则追

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

踪、人工复核和申诉反馈机制，才能形成可被平台、监管者与用户共同接受的治理闭环 [6, 12, 20-24, 28, 30]。

3.3 工业实践的可能路径

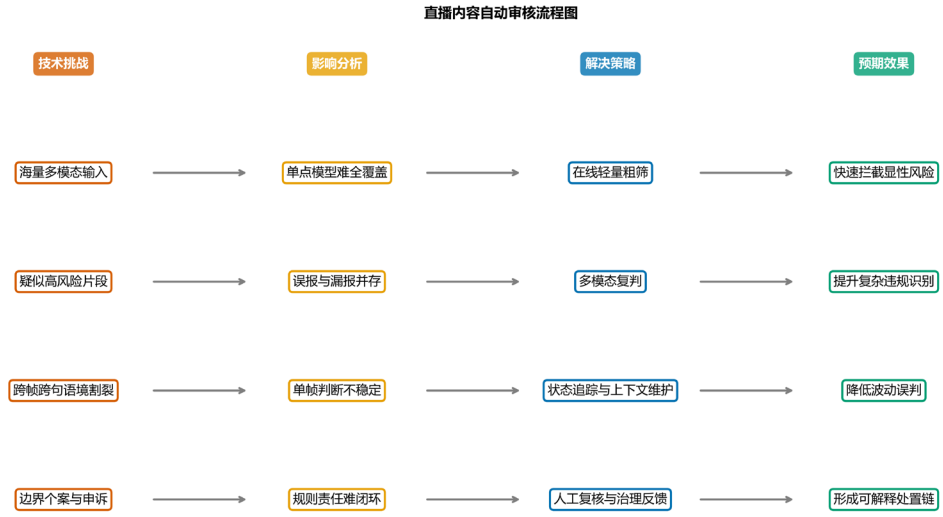


Figure 3. 直播内容自动审核流程图。

结合现有研究与平台实践，可以将较为可行的工业路径概括为三层。底层是高吞吐、低时延的在线过滤层，负责对文本流、关键帧、字幕与语音片段进行快速风险打分；中层是多模态复判层，聚焦于可疑样本的跨模态理解、证据补全与规则匹配；上层则是人工复核与治理层，负责处理低置信度样本、复杂边界个案和用户申诉 [2-3, 8, 13-18, 20-24, 29-30]。

这一架构的核心不在于寻找某一个“万能模型”，而在于通过任务拆分与资源分配，把不同模型部署在最适合的位置：轻量模型负责广覆盖，大模型负责深推理，人工审核承担最终裁量与责任闭环。只有当算法能力、系统架构与治理流程协同推进时，平台审核系统才可能在实际生产环境中兼顾规模、效率、合规性与用户可接受性。

4 结论与展望

4.1 主要结论

本文围绕直播与短视频平台内容自动审核技术的演进过程进行了系统梳理。总体而言，内容审核已经由基于规则和人工特征的离线批处理模式，发展为以深度学习、多模态融合和大模型增强为特征的复合式系统。离线方案仍在存量内容治理、精细复核与复杂证据整合方面具有重要价值；实时流式检测则成为直播场景中的关键基础能力；大模型正在推动审核系统由固定标签分类进一步走向规则理解与个案推理。

4.2 核心瓶颈

当前较为突出的瓶颈主要集中在四个方面：其一，复杂违规内容的跨模态理解仍不稳定；其二，低时延约束下的大模型部署成本仍然偏高；其三，新型风险样本稀缺导致模型持续学习与快速适配能力不足；其四，技术系统与治理规则之间仍缺乏充分稳定的接口，使审核可解释性、复核流程和

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

责任边界尚不清晰。

4.3 未来研究方向



Figure 4. 内容自动审核未来研究路线图。

未来研究可重点推进以下方向。第一，发展面向流式场景的轻量化多模态推理技术，包括自适应采样、分层缓存、蒸馏压缩与边云协同推理，以进一步缩短审核链路时延 [16-18, 27, 29]。第二，构建面向新型违规内容的持续学习与跨域迁移机制，降低模型对高成本人工标注的依赖，并提高对动态规则的适配能力 [15, 18-25]。第三，将平台政策、审核规则与模型决策过程进一步结构化，提升系统的可解释性、可追溯性和治理透明度 [12, 20-24, 28, 30]。第四，完善人机协同闭环，使自动审核系统不仅能够识别，而且能够说明、能够复核、能够申诉，从而更好支撑平台治理的长期稳定运行。

References

[1] ZHANG K X. Design and implementation of a short-text content moderation system based on large language models[D]. Beijing: Beijing University of Posts and Telecommunications, 2025.

[2] SONG L L. Construction and implementation of a multimodal intelligent moderation service platform based on multi-classification methods[J]. Radio & Television Network, 2025, 32(6): 90-93.

[3] LI C Y, LIU J, ZHANG J J. Research on a video content moderation platform based on BLIP and DeepSeek models[J]. Radio & Television Information, 2025, 32(12): 29-33.

[4] TIAN C L, WANG P, LIU S Y, et al. Research on AI-based video content moderation technologies[J]. Communication & Information Technology, 2025(Suppl.): 98-102.

[5] CHEN Y J, WANG T, XIE Y. Application of AI-assisted content review technologies in platform media content moderation[J]. Digital Technology & Application, 2025, 43(11): 81-83.

[6] NIU Y H. Research on key technologies for trustworthy collaborative media content moderation[D]. Beijing: Beijing Jiaotong University, 2024.

[7] AHMED S H, HU S, SUKTHANKAR G. The potential of vision-language models for content moderation of children's videos[C]//2023 22nd IEEE International Conference on Machine Learning and Applications. Piscataway, NJ: IEEE, 2023. DOI:10.1109/ICMLA58977.2023.00186.

[8] ZHANG W, SHI J, LIANG H, et al. Filter-and-refine: a MLLM-based cascade system for industrial-scale video content moderation[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Industry Track). Stroudsburg, PA: Association for Computational Linguistics, 2025: 1107-1118. DOI:10.18653/v1/2025.acl-industry.62.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

[9] CHEN F, YANG Z, ZHUANG B, et al. Streaming video diffusion: online video editing with diffusion models[C]//International Conference on Digital Image Computing: Techniques and Applications. Piscataway, NJ: IEEE, 2025. DOI:10.1109/DICTA68720.2025.11302466.
[10] OIKAWA Y, NAKAYAMA Y, MURAKAMI K. A stacking-based efficient method for toxic language detection on live streaming chat[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track. Stroudsburg, PA: Association for Computational Linguistics, 2022: 649-658. DOI:10.18653/v1/2022.emnlp-industry.58.
[11] FUJITAKE M, SUGIMOTO A. Real-time object detection by feature map forecast for live streaming video[C]//2021 IEEE International Conference on Multimedia and Expo. Piscataway, NJ: IEEE, 2021: 1-6. DOI:10.1109/ICME51207.2021.9428277.
[12] DOUEK E. Facebook's "Oversight Board": move fast with stable infrastructure and humility[J]. North Carolina Journal of Law & Technology, 2019, 21(1): 1-58.
[13] LIU X, LIU C, BAI R, et al. Offline detection of violations in Chinese e-commerce live streaming content[J]. IEEE Access, 2025, 13: 1-14. DOI:10.1109/ACCESS.2025.3584238.
[14] NGUYEN D V, LAM N H, NGUYEN A H, et al. MTikGuard system: a transformer-based multimodal system for child-safe content moderation on TikTok[EB/OL]. (2025-11-22)[2026-03-10]. https://arxiv.org/abs/2511.17955.
[15] LU X, ZHANG T, MENG C, et al. VLM as policy: common-law content moderation framework for short video platform[EB/OL]. (2025-04-21)[2026-03-10]. https://arxiv.org/abs/2504.14904.
[16] WANG H, JI D, ZHU L, et al. StreamSense: streaming social task detection with selective vision-language model routing[EB/OL]. (2026-01-30)[2026-03-10]. https://arxiv.org/abs/2601.22738.
[17] FREEMAN A C. Toward accessible and safe live streaming using distributed content filtering with MoQ[C]//2025 IEEE International Conference on Multimedia and Expo Workshops. Piscataway, NJ: IEEE, 2025: 1-7. DOI:10.1109/ICMEW68306.2025.11152153.
[18] LIANG H, SHI J, SHEN X, et al. Embedding-based retrieval in multimodal content moderation[EB/OL]. (2025-06-30)[2026-03-10]. https://arxiv.org/abs/2507.01066.
[19] CAO R. Online misinformation detection in live streaming videos[EB/OL]. (2025-03-16)[2026-03-10]. https://arxiv.org/abs/2503.12627.
[20] LEVI A, LEVI O, MISHRA S, et al. AI vs. human moderators: a comparative evaluation of multimodal LLMs in content moderation for brand safety[EB/OL]. (2025-08-07)[2026-03-10]. https://arxiv.org/abs/2508.05527.
[21] YADAV N, LIU J, ORTU F, et al. Revealing hidden mechanisms of cross-country content moderation with natural language processing[EB/OL]. (2025-03-07)[2026-03-10]. https://arxiv.org/abs/2503.05280.
[22] BLACKWELL L. Content moderation futures[EB/OL]. (2025-09-11)[2026-03-10]. https://arxiv.org/abs/2509.09076.
[23] PASCH S. AI vs. human judgment of content moderation: LLM-as-a-judge and ethics-based response refusals[EB/OL]. (2025-05-21)[2026-03-10]. https://arxiv.org/abs/2505.15365.
[24] GAO L, CHEN O, LEE R, et al. "I cannot write this because it violates our content policy": understanding content moderation policies and user experiences in generative AI products[C]//Proceedings of the 34th USENIX Security Symposium. Berkeley, CA: USENIX Association, 2025. https://www.usenix.org/conference/usenixsecurity25/presentation/gao-lan.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)

[25] FIROOZ H, LIU R, LU Y, et al. Scaling reinforcement learning for content moderation with large language models[EB/OL]. (2025-12-23)[2026-03-10]. https://arxiv.org/abs/2512.20061.
[26] GOCMEN C, LYKOURIS T, SINHA D, et al. Scheduling with uncertain holding costs and its application to content moderation[EB/OL]. (2025-05-27)[2026-03-10]. https://arxiv.org/abs/2505.21331.
[27] YAO L, LI Y, WEI Y, et al. TimeChat-Online: 80% visual tokens are naturally redundant in streaming videos[EB/OL]. (2025-04-24)[2026-03-10]. https://arxiv.org/abs/2504.17343.
[28] SHUKLA P, CHONG W Y, PATEL Y, et al. Silencing empowerment, allowing bigotry: auditing the moderation of hate speech on Twitch[EB/OL]. (2025-06-09)[2026-03-10]. https://arxiv.org/abs/2506.07667.
[29] CHEN J, WU Q, QIU X, et al. ModuRiskNet: automated multimodal content moderation and real-time risk detection framework for live streaming platforms[C]//2025 2nd International Conference on Software, Systems and Information Technology. Piscataway, NJ: IEEE, 2025. DOI:10.1109/SSITCON66133.2025.11341983.
[30] ELTAHER F, GAJULA R K, MIRALLES-PECHUÁN L, et al. Protecting young users on social media: evaluating the effectiveness of content moderation and legal safeguards on video sharing platforms[EB/OL]. (2025-05-16)[2026-03-10]. https://arxiv.org/abs/2505.11160.

© 2026 by the author(s); licensee Mason Publish Group, this work for open access publication is under the Creative Commons Attribution International License (CC BY 4.0). (http://creativecommons.org/licenses/by/4.0/)