SYSTEM AND METHOD FOR CONTENT SAFETY ENFORCEMENT DURING VIRTUAL VIDEO ASSEMBLY

TECHNICAL FIELD

[0001] The present invention relates generally to digital media processing, content moderation, and distributed video architectures. More particularly, the present invention relates to a system and method for classifying, flagging, and enforcing content safety policies at the individual, sub-file binary sample level during both media ingestion and dynamic, runtime virtual video assembly. The present invention further operates within an agentic content ecosystem and establishes a framework that shifts content safety from file-level analysis to deterministic and context-aware enforcement.

BACKGROUND OF THE INVENTION

[0002] Traditional digital video architectures rely on complete, monolithic files. In these environments, a video is authored, compressed, and wrapped inside a standard container format (such as MP4, MKV, or MOV) as a unified binary object.

[0003] Consequently, the entire discipline of digital content moderation and content safety has evolved around this monolithic file paradigm.

[0004] Every automated content safety system operates exclusively on complete, pre-assembled video files. When content is uploaded to a social media platform, video hosting service, or cloud storage repository, the platform ingests the complete file, passes the entire file through a series of centralized machine learning classifiers, and renders a binary or category-based moderation decision regarding the file as a visual and acoustic whole.

[0005] This approach to content moderation suffers from structural inefficiencies.

[0006] If a fraction of a two-hour video contains prohibited content, the entire file must be flagged, blocked, or removed. Conversely, to bypass these file-level filters, malicious actors frequently interleave prohibited content into long, innocuous video streams, overwhelming file-level ingestion pipelines and causing a high rate of false negatives or delayed compliance actions.

[0007] Furthermore, because legacy classification scans the file globally, it cannot naturally isolate or map the exact binary coordinates or sample positions responsible for the policy violation without running separate, computationally intensive post-processing passes.

[0008] The emergence of the agentic content world breaks the monolithic file paradigm entirely.

[0009] In an agentic ecosystem, autonomous artificial intelligence agents, automated workflows, and algorithmic systems dynamically compile, remix, and assemble video content on demand. These agents extract disparate fragments, frames, and audio-visual tracks from extensive source material libraries and stitch them together into customized streams tailored to individual viewers at the moment of consumption.

[0010] There is no pre-existing, static, complete video file to inspect, classify, or moderate prior to the moment a user initiates playback. Traditional content moderation systems cannot moderate something that does not exist structurally until the point of consumption.

[0011] Attempting to apply legacy content safety methods to agentic and dynamic video streaming creates security vulnerabilities and systemic failures. If a platform waits until an AI agent finishes assembling a dynamic video to run traditional file-level moderation classifiers, the classification pipeline introduces unacceptable latency, completely destroying the real-time interactivity required by generative media applications.

[0012] By the time the file-level moderation pipeline completes its analysis, the dynamically assembled stream has already been delivered through the network and consumed by the end-user.

[0013] Post-consumption moderation is a fundamental failure of safety infrastructure, as the psychological, social, or legal harm of exposing the viewer to unsafe or illegal content has already occurred.

[0014] Moreover, content safety is not a uniform or static global concept, it is inherently fluid, context-dependent, and multi-jurisdictional. What is legally permissible or socially acceptable in one geographic territory or country may be explicitly illegal or highly restricted in another.

[0015] Legacy content distribution networks and video platforms manage jurisdictional variance by duplicating files, maintaining separate localized regional buckets, and executing distinct, isolated moderation pipelines for each region. This duplication wastes exabytes of storage, creates massive synchronization overhead, and complicates regulatory auditing.

[0016] In a dynamic virtual assembly architecture, managing multi-jurisdictional compliance through file duplication is mathematically and operationally impossible. Because an AI agent can generate millions of unique, localized permutations of a video stream on the fly, a platform cannot pre-generate or pre-moderate every possible variation.

[0017] A content safety classification must be bound directly to the sub-file sample level and governed by a configurable, dynamic policy engine. This ensures that the exact same underlying source content can be safely assembled under vastly different safety regimes and legal frameworks for different jurisdictions, without duplicating physical files or running disjointed, parallel moderation pipelines.

[0018] Another structural limitation of legacy systems is the inability to detect and mitigate emergent safety concerns.

[0019] Traditional moderation tools analyze static frames or continuous audio files for explicit features like blood, weapon signatures, or explicit anatomy. However, they are fundamentally blind to context-driven safety violations that arise purely from the *sequence* or *combination* of individually safe media elements.

[0020] An AI agent could take an array of completely benign, clean, and certified video clips and arrange them in a specific order accompanied by automated text overlays that collectively constitute targeted harassment, severe defamation, or coordinated political disinformation. Because each individual source clip passes a content safety filter independently, legacy systems will clear the clips for ingestion, failing to recognize that the compound assembly forms an unsafe sequence.

[0021] Existing systems also decouple user access control, digital rights management (DRM), cryptographic provenance, and content safety into entirely separate, non-communicating operational silos. A media player might verify a DRM license token to grant playback access, but that token remains completely ignorant of whether the underlying samples contain content that violates local jurisdiction laws, platform terms of service, or user age restrictions.

[0022] This disconnected approach allows unverified accounts or underage users to bypass safety gates if the media player merely checks for file availability rather than assessing content safety constraints in tandem with user identity and regional attributes at the resolution layer.

[0023] Finally, compliance and mandatory reporting frameworks under current paradigms are highly retroactive, manually driven, and prone to platform-side manipulation or omission. When legacy systems discover highly illegal materials, such as CSAM, the process of documenting the exposure,

capturing the user metadata, identifying the exact digital signatures, and reporting the incident to regulatory bodies or law enforcement agencies (such as NCMEC or the IWF) involves asynchronous batch processing and voluntary reporting workflows.

[0024] There is a need for an automated system that embeds content safety classification directly into the structural identity manifest of sub-file samples and enforces compliance cryptographically at the point of resolution through tokenized assembly.

SUMMARY OF THE INVENTION

PROBLEMS TO BE SOLVED

[0025] It may be an advantage to develop a content safety framework that operates dynamically at the sub-file, individual binary sample level.

[0026] It may be an advantage to develop a system capable of moderating dynamically compiled video content within an agentic media ecosystem at the exact moment of virtual video assembly prior to user consumption.

[0027] It may be an advantage to develop a system that resolves multi-jurisdictional compliance variances for a single asset without the operational overhead of physical file duplication or running isolated, parallel regional moderation pipelines.

[0028] It may be an advantage to develop a system that maps precise content safety classifications, labels, and confidence scores directly to specific binary sample positions or temporal ranges

[0029] It may be an advantage to develop a method to detect and mitigate emergent safety concerns that arise purely from the specific sequence or combination of individually safe media elements.

[0030] It may be an advantage to develop a system that unifies user access control, cryptographic provenance, and content safety evaluation into a singular, interactive resolution layer.

[0031] It may be an advantage to develop a real-time policy enforcement engine capable of executing variable, granular responses , such as blocking, redacting, warning, or logging at the sub-file sample level during assembly.

[0032] It may be an advantage to develop an automated, immutable architectural reporting and auditing mechanism that triggers instantly upon interactions with mandatory compliance categories like child sexual abuse material (CSAM).

[0033] It may be an advantage to develop a verification method to prevent unauthorized or unsafe agentic AI video assemblies by validating proposed playback sequences against an immutable manifest before the video resolves.

MEANS FOR SOLVING THE PROBLEM

[0034] The present invention is directed to a system and method for classifying, flagging, and enforcing content safety policies at the individual binary sample level during both ingestion and virtual video assembly using a pluggable adapter and policy engine architecture.

[0035] The invention establishes a media virtualization pipeline where content safety classification becomes a native architectural property of every sub-file media sample, evaluated and enforced at the precise moment of assembly.

[0036] In an aspect of the present invention, the system for content safety classification and enforcement during virtual video assembly, comprise a media virtualization pipeline to index every sub-file media sample by its position during media ingestion; at least one pluggable content safety

classification adapter to evaluate inbound footage against specific content safety concerns and return a standardized result object; a policy engine to process a signed Content Safety Policy data object declaring context-specific safety requirements via Boolean operators; a registry module configured to generate a Content Safety Map (CSM) overlay upon classification that permanently records classification metrics for every flagged sample position to extend a Sample Identity Manifest (SIM) forming a compound SIM entry; a virtual video interface and a token-controlled assembly engine governed by a cryptographic video token encoding access control, licensing, authentication, and safety constraints; an automated auditing and reporting module embedded within a resolution layer to leverage verification events of the Video Token, SIM, and CSM to maintain an audit trail of assembly interactions with flagged content and compile compliance reports when a mandatory reporting category is triggered.

[0037] In another aspect of the present invention, a computer-implemented method for sample-level content safety classification, mapping, and enforcement during virtual video assembly, comprises the steps of ; indexing every sub-file media sample by specific position during ingestion within a virtual virtualization pipeline; passing inbound media through the content safety classification adapters to evaluate specific content safety concerns and return a standardized result object; declaring context-specific safety rules inside a signed content safety policy data object using Boolean operators; layering and stacking multiple safety policies to dynamically compute and enforce a single most restrictive combination of rulesets, producing a labelled classification record mapping specific samples to policies ; generating a Content Safety Map (CSM) overlay to extend a Sample Identity Manifest (SIM), forming a permanent compound SIM entry; generating a cryptographic video token for an assembly request, encoding content safety constraints alongside access control, licensing, and authentication criteria; requesting individual binary samples through a virtual interface at an exact moment of runtime virtual assembly and cross-referencing against the registered Content Safety Map in real time; executing a variable enforcement response determined by the Content Safety Policy, the

enforcement response selected from blocking the sample, redacting the sample by skipping flagged positions, delivering an automated content warning overlay, logging the access attempt, or initializing a reporting sequence; intercepting a proposed virtual video assembly sequence by an artificial intelligence agent from independent source files, cross-referencing independent source file Content Safety Maps, and refusing to validate the video token to block resolution if any sample violates the policy before the assembly resolves; executing a sequence-level classifier adapter to evaluate proposed sample orders to detect and intercept emergent combination safety concerns; and tracking validation events to maintain an unalterable complete audit trail and automatically generating and transmitting structured compliance reports containing unique identifiers, sample coordinates, requester identity, timestamp, and jurisdiction from a resolution layer for mandatory reporting categories.

[0038] Preferably, the standardized result object comprises a classifier identifier, a classification label, a confidence score, and a precise temporal range of flagged content expressed as specific binary sample positions.

[0039] Preferably, the pluggable adapters are selected from a group comprising a CSAM Detector, an Explicit Content Classifier, a Violence Classifier, a Hate Speech / Extremism Classifier, a Known Illegal Content Hash Registry Adapter, an Age-Restricted Content Classifier, and a Sequence-Level Classifier Adapter.

[0040] Preferably the CSAM Detector cross-references binary sample positions against established regulatory lists including PhotoDNA, NCMEC hash databases, and Internet Watch Foundation (IWF) registries.

[0041] Preferably, the explicit content classifier processes visual data to isolate nudity, sexual material, or explicit pornography under graduated labels such as suggestive, explicit, or extreme.

[0042] Preferably, the violence classifier identifies visual and auditory signatures of graphic violence, gore, or disturbing imagery, while the hate speech / extremism classifier parses embedded text tracks and audio streams to isolate hate speech or incitement.

[0043] Preferably, the known illegal content hash registry adapter functions directly at the binary layer to identify illicit material even if its visual presentation has been manipulated, and the age-restricted content classifier assigns traditional regional baseline ratings, such as G, PG, M, MA, and R equivalents, directly to the precise sub-file samples that trigger the restriction

[0044] Preferably, the policy engineis further configured to layer and stack multiple separate, independent safety policies from content owners, distribution platforms, and legal jurisdictions simultaneously to dynamically evaluate and enforce a single most restrictive combined ruleset.

[0045] Preferably, the compound SIM entry consists of a cryptographic sample identity layer, an authentication provenance layer, and a content safety classification layer anchored together to an immutable registry.

[0046] Preferably, the assembly engine is configured to cross-reference requested individual sample positions against the Content Safety Map in real time during virtual video assembly prior to resolution.

[0047] Preferably, content safety enforcement occurs strictly at the resolution layer, providing architectural prevention of policy violations before the media content can be published, delivered, or consumed.

[0048] Preferably, the content safety classification adapters return a standardized result object mapping safety labels, confidence scores, and classifier identifiers directly to specific temporal ranges expressed as sample positions rather than evaluating the file as a whole.

[0049] Preferably, the policy engine evaluates a Content Safety Policy structured as a signed data object containing Boolean logic strings including ALL OF, ANY OF, N OF M, and nested expressions.

[0050] The foregoing general description of the illustrative embodiments and the following detailed description thereof are merely exemplary aspects of the teachings of this disclosure and are not restrictive.

BRIEF DESCRIPTION OF THE FIGURES

[0051] Embodiments of the present disclosure will be discussed with reference to the accompanying figures wherein:

FIG. 1 illustrates a block diagram of the anatomy of a Compound Sample Identity Manifest (SIM) entry;

FIG. 2 illustrates a block diagram of the dual adapter architecture, showcasing the integration of forensic authentication adapters and content safety classification adapters into an evaluation framework;

FIG. 3 illustrates a sequential flow diagram contrasting conventional file-level video moderation with the granular, sample-level content safety classification and enforcement;

FIG. 4 is a data flow diagram illustrating the layered policy resolution framework;

FIG. 5 is a flow diagram illustrating an assembly-time content safety enforcement sequence governed by a cryptographic video token and a Content Safety Map lookup routine.;

FIG. 6 is a functional block diagram illustrating an agentic assembly safety routine;

FIG. 7 is a flow diagram illustrating an automated architectural reporting and audit trail ecosystem embedded natively within a media virtualization resolution layer; and

FIG. 8 is an architectural diagram illustrating a full platform stack layout.

DESCRIPTION OF THE INVENTION

[0052] The reference to any prior art in this specification is not, and should not be taken as, an acknowledgement or any form of suggestion that such prior art forms part of the common general knowledge.

[0053] It will be understood that the terms “comprise” and “include” and any of their derivatives (e.g. comprises, comprising, includes, including) as used in this specification, and the claims that follow, is to be taken to be inclusive of features to which the term refers, and is not meant to exclude the presence of any additional features unless otherwise stated or implied.

[0054] In some cases, a single embodiment may, for succinctness and/or to assist in understanding the scope of the disclosure, combine multiple features. It is to be understood that in such a case, these multiple features may be provided separately (in separate embodiments), or in any other suitable combination. Alternatively, where separate features are described in separate embodiments, these separate features may be combined into a single embodiment unless otherwise stated or implied. This also applies to the claims which can be recombined in any combination. That is a claim may be amended to include a feature defined in any other claim. Further a phrase referring to “at least one of” a list of items refers to any combination of those items, including single members. As an example, “at least one of: a, b, or c” is intended to cover: a, b, c, a-b, a-c, b-c, and a-b-c.

[0055] The term Sample Identity Manifest (SIM) refers to a compound cryptographic record encapsulating both Sample Identity Data (such as binary hashes) and an Authentication Provenance Bundle (the verification audit record).

[0056] The term authentication policy engine means a configurable rules layer that acts as a gatekeeper for SIM registration. It evaluates whether an asset passes verification metrics and manages the conditions under which an asset can be formally added to an immutable ledger or registry.

[0057] The term authentication source adapter refers to a discrete, pluggable software module designed to evaluate media against a specific forensic or cryptographic standard.

[0058] The term Content Safety Map (CSM) refers to a sample-level data overlay that extends, rather than replaces, the Sample Identity Manifest (SIM).

[0059] The term video token refers to a cryptographically validated control layer generated for an assembly request that governs assembly and media resolution conditions.

[0060] The term immutable registry means a timestamped data store implemented via a distributed ledger or blockchain architecture.

[0061] The present invention is directed to a system and method for classifying, flagging, and enforcing content safety policies at the individual binary sample level during both ingestion and virtual video assembly. By utilizing content safety markers to govern the assembly of the virtual container, the invention shifts content moderation from an external, post-processing file operation into an intrinsic property of the underlying media elements themselves.

[0062] This shift addresses the critical shortcomings of traditional content safety solutions, which operate globally on pre-assembled video files and cannot protect users in dynamic, real-time media environments.

[0063] Within an agentic content ecosystem, where artificial intelligence agents dynamically compile and assemble video tracks from disparate fragments of source material, complete media files do not exist prior to the exact point of consumption. Traditional video moderation platforms cannot evaluate something that does not exist structurally until a user initiates playback, resulting in severe latency or post-consumption exposure harms.

[0064] The present invention solves this problem through a virtual video architecture where media content only assembles at the precise point of resolution through a specialized virtual interface.

[0065] Additionally, the present invention accommodates fluid, context-dependent jurisdictional requirements where media items legal in one territory may be strictly prohibited in another.

[0066] The system includes a media virtualization pipeline configured to index every sub-file media sample by its exact binary coordinate and playback position during asset ingestion. Because every audio and video fragment is precisely indexed from the moment it enters the pipeline, downstream evaluation modules can isolate specific sub-file parameters rather than rendering a global decision on the file as a whole.

[0067] This tracking structure serves as the computational baseline for all sample-level tracking, mapping, and real-time execution layers within the platform.

[0068] The system further includes content safety classification adapters, integrated into the media virtualization pipeline alongside baseline authentication adapters.

[0069] Each content safety adapter functions as a discrete processing module that evaluates inbound footage streams against a specific content safety concern. Upon completing an evaluation scan, the adapter returns a highly structured, standardized result object containing a unique classifier identifier, an explicit classification label, a confidence score, and a precise temporal range of flagged content expressed as specific binary sample positions.

[0070] The system implements a CSAM Detector adapter configured to evaluate visual media streams against trained classifiers for child sexual abuse material. To ensure compliance with mandatory international tracking registries, the CSAM Detector cross-references binary sample signatures against established law enforcement and regulatory hash databases, including PhotoDNA, NCMEC hash lists, and Internet Watch Foundation (IWF) registries.

[0071] When an infraction is caught, the adapter returns the exact sample-level positions of the flagged content to enable instantaneous sub-file isolation. Crucially, an operational distinction is maintained between standard detection and the system's unique addressing layer. While pluggable modules evaluate inbound footage through traditional artificial intelligence, technology-based visual recognition, or against existing records by SMPTE timecode, the binary sample positions themselves are only ever made accessible through the Sample Identity Manifest (SIM) or the indexed file generated by the process.

[0072] An explicit content classifier adapter is provided within the system to systematically analyze inbound visual content for nudity, sexual acts, and pornographic material. This module applies graduated severity labels, such as suggestive, explicit, or extreme, based on trained neural classification thresholds. The classifier maps these assigned labels directly to specific sample positions, ensuring that benign portions of a video asset remain decoupled from isolated explicit frames.

[0073] The system further incorporates a violence classifier adapter designed to parse both visual frames and accompanying acoustic data streams for graphic violence, weapons signatures, gore, and disturbing imagery. This adapter assigns graduated severity metrics to the processed footage and records the exact sample-position coordinates where the violent imagery or sounds occur. This allows the platform to pinpoint intense structural sequences within an extensive media asset without altering the surrounding video material.

[0074] A hate speech/extremism classifier adapter is integrated within the system to analyze decoded audio tracks and any embedded text or subtitle arrays for hate speech, extremist propaganda, or incitement to violence. By parsing acoustic streams and text overlays in tandem, this module detects verbal or text-based policy violations and maps the resulting infraction labels directly to the corresponding audio and video sample positions.

[0075] A known illegal content hash registry adapter functions directly at the binary layer of the media virtualization pipeline rather than processing visual features. This adapter cross-references binary sample hashes against international registries of known illegal content maintained by law enforcement and global regulatory bodies. Because it operates at the binary level, the module can reliably detect known illegal material even if the visual presentation has been cropped, filtered, or otherwise altered to bypass standard visual classifiers.

[0076] The system additionally features an age-restricted content classifier adapter configured to evaluate media tracks against traditional regional rating frameworks, such as G, PG, M, MA, and R equivalents. Rather than assigning a blanket rating to an undivided file, this classifier generates a sample-level mapping that flags the exact sub-file samples or scenes that trigger each respective rating tier. This allows downstream assembly tools to al

[0077] ter content matching a user's age criteria without switching to a different file.

[0078] The system integrates a sequence-level classifier adapter that evaluates the proposed assembly layout of a virtual video as an ordered whole, rather than assessing individual source samples in isolation, to detect and block emergent safety concerns. This adapter identifies scenarios where individually benign, clean clips are ordered maliciously by an agent to form compound sequences that collectively constitute targeted harassment, defamation, or political disinformation.

[0079] Operating with the classification adapters is a content safety policy engine that processes a signed content safety policy data object that declares context-specific safety requirements for given viewing environments.

[0080] The policy engine utilizes Boolean evaluation logic string, including ALL OF, ANY OF, N OF M, and multi-tier structural nesting to define permissible media boundaries. This engine acts as a dynamic rules layer that evaluates whether an asset's sub-file components comply with contextual platform constraints.

[0081] The policy engine is structurally configured to layer and stack multiple separate, independent safety policies simultaneously, including a content owner's policy at ingestion, a platform's delivery rules at distribution, and a local territory's legal requirements at the point of assembly.

[0082] The engine treats these layered criteria as a cumulative stack rather than mutually exclusive choices. It dynamically computes all active rules and enforces the single most restrictive combination of the stacked rulesets across the requested viewing session.

[0083] A registry module is coupled to the policy engine to generate a Content Safety Map (CSM) overlay upon asset classification. The CSM operates as a secure data overlay that permanently records the classification label, confidence score, classifier identifier, and governing policy parameters for every single flagged sample position.

[0084] The registry module applies the CSM as a structural extension layer that extends, rather than replaces, the underlying Sample Identity Manifest (SIM).

[0085] The registry module compiles these elements into a compound SIM entry consisting of three permanent, integrated architectural layers: a first sample identity layer containing cryptographic hashes of individual binary samples; a second authentication provenance layer capturing human-creation verification records; and a third content safety classification layer embedding the sample-level CSM. This three-layer manifest configuration ensures that an asset's safety metrics are permanently tied to its cryptographic identity.

[0086] The system connects this compound manifest structure to an immutable registry implemented via a distributed ledger. The registry module hashes all three manifest layers together and anchors them as a single unified entry to the ledger. This immutable anchor point ensures that content safety maps cannot be altered, bypassed, or deleted by platform operators, unauthorized users, or external tampering attempts.

[0087] To govern media resolution at the point of consumption, the system features a token-controlled assembly engine linked to a virtual video interface. The assembly engine is governed by a cryptographic Video Token generated for an assembly request, which encodes precise runtime constraints including user access control, licensing boundaries, and stacked safety criteria. As individual binary samples are pulled, the engine cross-references their coordinate positions against the registered CSM in real time to enforce compliance before any frames resolve on a user's display.

[0088] Finally, the system includes an automated auditing and reporting module embedded directly within the structural resolution layer itself. This module tracks all token generations, manifest verifications, and assembly requests to maintain an unalterable, complete audit trail of user and agent interactions with flagged materials.

[0089] For severe safety violations falling into mandatory compliance categories, the module automatically compiles a structured compliance report and dispatches it directly to law enforcement and regulatory authorities.

[0090] FIG. 1 illustrates the structural anatomy 100 of a registered compound Sample Identity Manifest (SIM) entry 101 anchored to an Immutable Registry 102.

[0091] The compound manifest entry 101 represents an incremental, multi-layered data structure that merges asset identity, verified provenance, and sample-mapped safety coordinates into a singular root record.

[0092] The boundaries of the compound entry 101 encapsulate three distinct operational zones, designated as Layer 1 103, Layer 2 104, and Layer 3 105 which are compiled during ingestion and permanently committed together under a single, non-repudiable compound root hash string.

[0093] Layer 1 defines the sample identity base layer 103 carried over from the core capabilities of right to identify references. Layer 1 103 contains an ordered series of independent cryptographic hash pills (e.g., Hash: a7f3... 106, Hash: c9b2...107, and Hash: f1e4...108) that correspond to the raw binary composition of individual media samples processed through the virtualization pipeline. These individual hash pills are combined via a structural tree configuration to yield a localized root hash value 109 (e.g., 0xR1).

[0094] Additionally, Layer 1 103 encodes the parameters of the selection algorithm 110 utilized during ingestion, such as an I-frame biased selection rule block, ensuring absolute alignment between the physical file structure and the manifest layout.

[0095] Layer 2 defines the authentication provenance layer 104, which maintains the unalterable forensic validation profile of the ingested asset as established by right to identify references. Layer 2

104contains a governing profile reference 111(e.g., Policy: Studio_Feature) and hosts an array of cryptographic and forensic verification indicator fields.

[0096] These fields track independent verification modules and record affirmative status checkmarks, including a C2PA Provenance Verifier field 112 (C2PA: PASS), a camera sensor forensic field 113 (PRNU: PASS), and a deepfake classification field 114 (AI Detector: NEGATIVE), ensuring the asset is certified as authentic and human-created before safety properties are applied.

[0097] Layer 3 105 establishes the content safety classification layer, which introduces the core structural extension of the present invention.

[0098] Layer 3 105 implements a comprehensive Sample-Position Safety Map 115 that charts safety attributes across an indexed timeline range (e.g., samples 0 -> 6000). The timeline is divided into discrete, coordinate-mapped blocks comprising clean spans and explicitly flagged violation ranges.

[0099] As shown, a first violating coordinate range (mapping samples 4217-4892) is explicitly flagged under an explicit content label 117 and bound to metadata blocks tracking a specific module ID (classifier: CSC-EXP-2), an active logic source 120 (policy: Jurisdiction_AU), a statistical reliability score (conf 0.97) 122, and a mandated enforcement mechanism 121 (enforce: AGE-GATE).

[00100] A secondary violating coordinate range (mapping samples 5100-5240) is similarly flagged under a violence label 118 and bound to accompanying metadata fields capturing an alternative module ID 123 (classifier: CSC-VIO-1), an alternative tracking source 124 (policy: Platform_Std), an alternate statistical threshold (conf 0.88) 125, and an alternate enforcement path (enforce: WARN) 126.

[00101] FIG. 2 illustrates the functional layout of the dual adapter architecture 200, demonstrating how the interface structures of the right to identify references are extended to form a general-purpose evaluation framework.

[00102] Inbound footage 201 is split, mapped, and systematically indexed by sample position during initial ingestion, generating a steady stream of discrete binary blocks. These blocks are delivered directly to a unified pluggable adapter interface 202 that functions as a single physical socket layer situated directly between the inbound footage 201 and the adapter banks 203 and 204 capable of hosting multiple distinct categories of processing adapters simultaneously.

[00103] The pluggable adapter interface hosts two independent evaluation families that process inbound sample chunks in parallel: an authentication source adapters family 203 and a content safety classification adapters family 204.

[00104] The authentication source adapters family 203 contains modular verification components engineered to resolve asset integrity questions (answering: "Is it real?"), comprising a C2PA Provenance Verifier module 205, a PRNU Sensor Match module 206, and an AI / Deepfake Detector module 207.

[00105] The content safety classification adapters family 204 contains independent protective modules engineered to parse semantic and material boundaries (answering: "is it safe?"), comprising a CSAM Detector module 208 (utilizing PhotoDNA, NCMEC, and IWF registry logic), an explicit content classifier module 209, a violence classifier module 210, a hate speech / extremism classifier module 211, a known illegal content hash registry module 212, and an age-restricted content classifier module 213.

[00106] Each adapter docking with the pluggable adapter interface 202 outputs data conforming to a standardized result object 214 shape. This standardized object 214 configuration enforces uniform compliance tracking across different software modules, encapsulating a classifier field 215 (e.g., "CSC-EXP-2"), a category label field 216 (e.g., "EXPLICIT"), a numerical accuracy confidence field 217 (e.g., 0.97), and a precise sub-file coordinates field designated as sample range 218 (e.g., 4217 - 4892). Because the standardized result object isolates violations by sample positions rather than the whole file, it can feed precise structural data directly into a shared policy engine 219.

[00107] The policy engine 219 evaluates these incoming objects using complex, nested Boolean logic configurations (including ALL OF 220, ANY OF 221, and N OF M 222strings) to render a registration decision, subsequently gating asset entry directly into the final Compound SIM container 223.

[00108] FIG. 3 provides a comparative visualization of a conventional file-level moderation paradigm 301 and the granular sample-level classification paradigm 302 of the present invention.

[00109] Under the conventional file-level moderation 301 , traditional safety classifiers 303 analyze an uploaded media file as an undivided monolithic block. If any explicit content is detected anywhere within the timeline, the global classifier issues a single un-isolated verdict block 304 stating that explicit content is present throughout the asset. This prompts the platform to trigger a blanket blocked indicator status across the entire standalone video.mp4 container file, resulting in an absolute file rejection output 305 that removes the clean majority of the video frames along with the offending sequence because no partial outcome is technically possible.

[00110] In contrast, the sample-level classification 302 indexes an ION virtual video track by individual sample positions to enable isolated processing. A granular sample scanner evaluates the

asset and restricts its infraction findings to a precise sub-file coordinate span 306 (stating: "samples 4217-4892 = EXPLICIT") with an associated validation score.

[00111] This generates a mapped classified sample stream 307 across the asset duration (e.g., blocks 0 through 6000), cleanly separating unflagged baseline blocks from an isolated flagged block range.

[00112] During downstream delivery requests, a policy engine processes this map to generate a customized, context-aware policy-driven assembly outcome layer 308. Clean sample blocks coordinate with a clear RESOLVE 309 instruction, whereas the flagged block range triggers a targeted BLOCK / REDACT 310 command on the fly. As a result, a dynamic resolution engine can deliver a tailored playback outcome 311 where the clean spans still resolve perfectly (e.g., resolving 17 / 20 clean sample spans while keeping 3 offending samples completely redacted).

[00113] FIG. 4 illustrates the operational mechanics of the layered policy resolution framework 400, demonstrating how separate independent compliance criteria stack together to dictate effective real-time assembly behaviors.

[00114] The framework 400 processes a cumulative contextual policy stack that applies to a single virtual video assembly, containing a content owner studio policy block 401 checked during ingestion, a distribution platform rules block 402 verified during syndication, and a localized territory jurisdiction policy block 403 evaluated at the exact point of consumption.

[00115] Rather than treating these entities as mutually exclusive alternatives, a multi-category matrix evaluator 424 processes their criteria concurrently, applying an absolute mathematical severity gradation scale (ranking restriction modes from least to most severe: ALLOW < WARN < AGE-GATE < REDACT < BLOCK < BLOCK+REPORT) to guarantee that the single most restrictive applicable rule always wins each respective content category.

[00116] Across an explicit content evaluation category 404, the content owner policy block mandates an ALLOW 408 threshold, the platform rules block requires an AGE-GATE 409 constraint, and the local jurisdiction policy block demands an AGE-GATE 410 tracking line. The stacking logic resolves these values against the severity gradation scale, determining that the matching AGE-GATE restriction represents the most restrictive active value and outputting an effective combined policy command of AGE-GATE 420 derived from the platform and jurisdiction parameters.

[00117] Furthermore, across a graphic violence evaluation category 405, the content owner policy block indicates an ALLOW 41 tier, the platform rules block enforces a WARN 412 overlay requirement, and the local jurisdiction policy block returns an ALLOW 413 baseline.

[00118] Comparing these settings against the scale isolates the platform-driven WARN action as the most restrictive configuration, establishing an effective combined policy output of WARN 421 across the active viewing session.

[00119] Across a Hate Speech evaluation category 406 , all three independent policy structures—content owner, platform, and jurisdiction, unanimously mandate an absolute BLOCK 414, 415, 416 restriction. Because all three blocks agree, the system bypasses comparative sorting and outputs an absolute effective policy decision of BLOCK 422 to structurally suppress the offending sample ranges.

[00120] Across a CSAM evaluation category 407, the content owner policy block, the platform rules block, and the local jurisdiction policy block all simultaneously mandate a maximum-severity BLOCK+REPORT enforcement action 417, 418, 419. This compliance alignment instructs the stacking framework to immediately issue a mandatory effective policy output of BLOCK+REPORT 423. The framework demonstrates that because layering is executed per individual content category, a single un-duplicated source asset can be safely distributed under vastly different safety regimes for

different territories without requiring physical file duplication or disjointed regional moderation workflows.

[00121] FIG. 5 illustrates the sequential operational flow of an assembly-time content safety enforcement routine 500.

[00122] The routine 500 initializes when an automated AI agent or a human end-user dispatches an assembly request vector 501 into the platform. The system intercepts the request and generates a contextual, cryptographic video token 502, encoding distinct metadata security layers comprising user access control credentials 503 , digital rights management licensing bounds 504, origin authentication profiles 505, and context-specific safety constraints 506.

[00123] The validated video token 502 is delivered to a virtual video interface engine 507 that initiates a loop requesting a reference to the individual media samples from the virtual video index, to reconstruct a new virtual container. As each sample coordinate is called, the system triggers an assembly-time safety check gate 512. The check gate 512 aggregates active consumer context inputs (including localized user jurisdiction 509, platform rules 510, and verified age-verification status metrics 511) and initiates a real-time Content Safety Map lookup channel 508 to evaluate the targeted sample coordinate against the manifest records anchored to the underlying SIM.

[00124] If the check gate 512 returns a negative compliance violation result 513 (NO), indicating that the requested coordinate is clean, the routine routes along an unflagged resolution path to execute a RESOLVE 514 bmodule that appends the sample block to the active assembly stream, subsequently updating a tracking counter to request the next sample position (N+1) 515.

[00125] If the check gate 512 returns an affirmative violation result 516 (YES, flagged & disallowed), the routine branches along a disallowed path to engage a dynamic enforcement selector array 517 where a policy-driven enforcement mode fires based on the active policy context.

[00126] The selector array contains multiple distinct functional execution modules comprising a Block module 518 (configured to refuse sample resolution and terminate the stream or substitute a static placeholder), a Redact module 519 (configured to seamlessly skip the flagged coordinates and compile the remaining clean frames on the fly), a Warn module 520 (configured to deliver the media accompanied by an automated warning overlay or container flag), a Log module 521 (configured to record the access attempt within an unalterable registry ledger), and a Report module522 (configured to auto-report the violation payload to designated authorities).

[00127] The system matches safety labels to specific operational paths; a verified CSAM detection automatically fires a mandatory combined BLOCK + LOG + REPORT sequence 523 whereas explicit material requests from unverified accounts trigger an AGE-GATE / WARN response 524 , and graphic violence ranges trigger a targeted WARN overlay 525.

[00128] FIG. 6 illustrates the operational mechanics of the agentic assembly safety architecture 600, highlighting the dual-stage verification sequence executed to protect autonomous workflows within generative media environments.

[00129] The architecture 600 processes environment interactions where an automated AI Agent entity 608 selects discrete sample fragments across multiple disparate source file libraries 601 , as Source A 602, Source B 603 , and Source C 604, each carrying its own independent manifest tracking layers comprising local SIM and CSM data blocks 605, 606, 607. The AI Agent entity 608 aggregates these fragments into a proposed assembly layout candidate 609 representing the ordered playback sequence the agent intends to resolve for an end-user.

[00130] Before the requested layout candidate is permitted to play back, the system intercepts the stream and routes the candidate through a Pre-Resolution Evaluation framework 610 to achieve absolute architectural prevention before delivery or consumption occurs.

[00131] The evaluation framework 610 executes two distinct compliance check sequences: a per-sample classification routine 612 and a holistic sequence-level classifier routine 611 .

[00132] The per-sample classification routine 612 extracts each individual selected sample chunk and validates its binary coordinates against its respective source manifest layer (checking the sample against the source CSM and active policy stack) to verify that all individual components return an ALL PASS 613 safety status.

[00133] The sequence-level classifier 611 routine parses the ordered layout of the proposed sequence as a single combined whole to track and flag emergent harm indicators 614. The sequence-level classifier 611 is engineered to identify context-driven risks where individually benign, clean clips 618 are arranged in a malicious layout order that collectively forms targeted harassment, severe defamation, or political disinformation.

[00134] If the sequence classifier flags an emergent risk 616 , the system routes data to a structural termination node that forces a TOKEN DOES NOT VALIDATE 617 status, completely blocking the assembly from resolving or playing back, or alternatively, removing the offending blocks and reinstigating reassembly of a new container minus the offending blocks.

[00135] FIG. 7 illustrates the functional data flow of the automated architectural reporting and audit trail ecosystem 700 for automated compliance tracking.

[00136] The ecosystem 700 is activated at a target assembly event trigger point when a user or agent assembly request involves a sub-file sample coordinate falling within a critical mandatory-reporting category (e.g., a verified child sexual abuse material detection).

[00137] Because the assembly session 701 is continuously governed by a cryptographic video token, verified against a sample identity manifest, and checked against a Content Safety Map, the resolution layer automatically triggers702 an Immutable Structured Report constructor block .

[00138] The constructor block 703 automatically extracts real-time operational tracking metrics directly from the resolution event, populating an array of immutable data fields comprising a unique asset identifier 704 (SIM identifier: 0x4b9c...9e21), the exact offending coordinates 705 (Flagged sample positions: 4217 - 4892), the processing module signature 706 (Classifier result: CSAM · conf 0.99 · CSC-CSAM-1), the verified consumer identity string 707 (Requesting party: token holder #A1932), a non-repudiable system clock value 708 (Timestamp: 2026-06-01T09:43Z), and the request origin territory 709 (Jurisdiction: AU).

[00139] The constructor block 703 further splits this populated compliance payload across two parallel, automated data paths to ensure regulatory tracking without relying on voluntary platform administration.

[00140] A first automated path 710 routes along a non-voluntary reporting dispatch line to transmit the structured report payload directly to an external law enforcement or regulatory body receiver node with full cryptographic provenance chains 713 attached.

[00141] A second automated path 711 routes data into a tamper-evident audit trail logger that records every user interaction, token validation event, and cross-reference block associated with the flagged content.

[00142] The logger cryptographically hashes the audit tracking log and anchors it directly to the target immutable registry ledger 712 providing an unalterable, attributable verification record built into the resolution layer itself.

[00143] FIG. 8 refers to a block diagram of the multi-layered platform architecture stack 800, illustrating how the present system nests natively within the underlying functional layers of the media virtualization platform.

[00144] The stack 800 organizes system dependencies across five distinct application tiers that build sequentially from the bottom up along a series of inter-layer dependency vectors to combine system operations into a unified workflow: virtualize, record, govern, authenticate, and protect.

[00145] At the baseline of the stack , Layer 1 801 defines the structural foundation (LAYER 1 VIRTUALISE), which utilizes the core parameters to implement virtual video containers 802 and establish sample-level binary addressing across media streams.

[00146] Directly above this foundation, Layer 2 803 establishes the cryptographic recording tier (LAYER 2 RECORD), which implements a distributed ledger to host an immutable, secure registry of every committed manifest entry across downstream delivery loops.

[00147] Layer 3 805 defines the platform governance tier (LAYER 3 GOVERN), which utilizes the core capabilities to manage a token-controlled assembly layer that dynamically compiles media fragments based on validated video tokens 806.

[00148] Layer 4 807 establishes the forensic verification tier (LAYER 4 AUTHENTICATE), which implements the foundational provenance engine of right to identify references to evaluate forensic standard adapters, verify camera metadata, and prove media content 808 is genuine.

[00149] Layer 5 809 defines the protective application layer added by the present invention (LAYER 5 PROTECT). Layer 5 809 overlays sample-level content safety classification and context-aware policy enforcement 810 across the entire stack, working in tandem with the lower authentication, governance, recording, and structural addressing layers 811 to deliver deterministic safety containment at the exact moment of media assembly.

[00150] The computer-implemented method for sample-level content safety classification, mapping, and enforcement during virtual video assembly executes across a series of definitive, sequential stages.

[00151] The method begins with an ingestion step where inbound media footage enters the virtualization pipeline. During this ingestion stage, the media virtualization pipeline processes the raw asset and systematically indexes every sub-file media sample by its exact binary coordinate and playback position to establish a granular coordinate baseline.

[00152] The method proceeds by passing the inbound media through one or more of the pluggable Content Safety Classification Adapters integrated within the virtualization pipeline. The adapters execute parallel visual, auditory, and binary scans to evaluate the media against specific content safety concerns, including child sexual abuse material, explicit content, violence, hate speech, known illegal content hashes, and age restrictions. Each active adapter processes the sub-file media stream and generates an independent safety assessment.

[00153] Each adapter then formulates and returns a standardized result object containing its classifier identifier, an assigned classification label, a confidence score, and a precise temporal range of flagged frames.

[00154] The method maps these returned classification parameters directly to specific binary sample positions or block coordinates rather than applying the moderation decision globally to the entire file. This isolates any violating frames while verifying that surrounding, non-violating content remains unflagged.

[00155] Simultaneously, the method declares context-specific safety regulations inside a signed Content Safety Policy data object. This policy data object specifies precise evaluation rules using nested Boolean logic strings such as ALL OF, ANY OF, and N OF M. This step establishes the structural criteria against which the indexed sample coordinates will be validated during downstream delivery and assembly requests.

[00156] The policy engine then executes a policy stacking step where it layers and stacks multiple independent safety policies from separate entities, including the content owner's rules at ingestion, the platform's delivery guidelines at distribution, and a local territory's legal requirements at assembly.

[00157] The method further evaluates these stacked policies concurrently to dynamically calculate and enforce the single most restrictive combination of rulesets across a given user session. This produces a multi-state classification record mapping specific samples to policies, allowing asset registration with an accompanying safety map instead of issuing a binary pass/fail rejection.

[00158] The method then generates a Content Safety Map (CSM) overlay that permanently records these classification labels, scores, and governing policy metrics for every flagged sample position. The registry module applies this CSM overlay directly onto the asset's existing Sample Identity Manifest (SIM), extending the core manifest rather than replacing it. This step forms a permanent, compound SIM entry that securely couples the asset's safety profile with its cryptographic baseline.

[00159] To secure this data, the method organizes the compound SIM entry into three distinct, permanent layers: the Sample Identity cryptographic hash layer, the Authentication Provenance layer, and the Content Safety Classification layer. The registry module collectively hashes all three layers together and anchors the unified cryptographic payload directly to an immutable ledger registry. This anchoring step guarantees absolute data permanence and non-repudiation across all downstream platforms.

[00160] When a user or automated system requests a virtual video assembly, the virtual video interface catches the request and generates a contextual cryptographic Video Token. The system encodes precise content safety constraints, access controls, licensing boundaries, and regional jurisdictional requirements directly inside the generated Video Token. This token acts as the primary governance layer that dictates assembly conditions during the active playback session.

[00161] At token validation, the virtual video interface evaluates the proposed assembly manifest against the Content Safety Map. All verification happens before the final empty container is delivered, ensuring the right to resolve by checking the Content Token against the Content Safety Map.

[00162] For a novel sequence an agent proposes across multiple files, the proposed layout is evaluated against each source's Content Safety Map before the token validates. This guarantees compliance verification is completed prior to the token resolving, structurally preventing the assembly from resolving or removing the offending blocks to reinstigate reassembly of a new container if any sample violates the policy.

[00163] At the runtime moment of virtual video assembly, the virtual video interface requests individual binary samples from storage layers to compile the video on the fly. As each sample is pulled, the assembly engine cross-references its precise position coordinate directly against the

anchored Content Safety Map in real time. This step ensures that every frame is checked for compliance prior to resolution and playback.

[00164] If a requested sample coordinate matches a flagged entry that violates the active policy context encoded in the Video Token, the assembly engine dynamically determines and executes a variable enforcement response mandated by the policy. Rather than relying on a fixed platform response, the method selects an enforcement mode dynamically from a group comprising a Block mode, a Redact mode, a Warn mode, a Log mode, and a Report mode.

[00165] Depending on the calculated severity, the engine executes a Block mode to refuse sample resolution and terminate the stream or substitute a static placeholder. Alternatively, it executes a Redact mode to seamlessly skip flagged coordinates and compile the remaining clean samples on the fly without modifying the underlying source files. It can also execute a Warn mode to deliver the content with an automated warning overlay, or a Log mode to record the access attempt for compliance auditing.

[00166] To secure autonomous workflows within generative media applications, the method actively intercepts proposed virtual video assembly sequences compiled by artificial intelligence agents extracting fragments across multiple independent source files.

[00167] The system retrieves the independent SIM entry and associated Content Safety Map for each disparate source asset utilized by the AI agent and checks the combined layout against active safety policies prior to delivery. If any sample within the agent's proposed assembly sequence violates the target policy context, the cryptographic Video Token fails validation, and the system structurally prevents the assembly from resolving or playing back.

[00168] Concurrently, the method executes the sequence-level classifier adapter to evaluate the proposed playback layout as an ordered whole. This step detects and blocks emergent safety concerns where individually benign, clean clips are ordered maliciously by an agent to form a sequence that collectively constitutes harassment or defamation. If an emergent risk is detected, the assembly token is invalidated, preventing the assembly from resolving.

[00169] Finally, the method tracks all token validation and manifest verification events to maintain an unalterable complete audit trail of user and agent interactions with flagged materials.

[00170] For severe safety violations falling into mandatory compliance categories, such as verified detections of child sexual abuse material (CSAM), the module automatically triggers a non-voluntary reporting sequence handled natively at the resolution layer.

[00171] The system instantly compiles a structured compliance report populated with the unique SIM identifier, the exact sample positions flagged, the specific classifier results, the identity of the requesting party retrieved from the validated Video Token, the precise timestamp, and the requesting jurisdiction, and transmits the payload directly to law enforcement and regulatory authorities.

[00172] It will be appreciated by those skilled in the art that the disclosure is not restricted in its use to the particular application or applications described. Neither is the present disclosure restricted in its preferred embodiment with regard to the particular elements and/or features described or depicted herein. It will be appreciated that the disclosure is not limited to the embodiment or embodiments disclosed, but is capable of numerous rearrangements, modifications and substitutions without departing from the scope as set forth and defined by the following claims.

[00173] Please note that the following claims are provisional claims only, and are provided as examples of possible claims and are not intended to limit the scope of what may be claimed in any

future patent applications based on the present application. Integers may be added to or omitted from the example claims at a later date so as to further define or re-define the scope of the invention.