Autonomous Personalized Cinema: A Narrative Review and Conceptual Framework

Artificial intelligence is increasingly used in screenplay development, visual production, editing, sound generation, audience analysis, recommendation, and media personalization. These applications, however, remain fragmented and largely dependent on human-directed workflows. This narrative review examines the technical, creative, cognitive, evaluative, and ethical conditions under which an artificial-intelligence system could achieve runtime creative autonomy and generate a complete film for an individual viewer. The study synthesizes research across computational creativity, long-form narrative generation, multimodal media synthesis, multi-agent production, viewer modeling, digital twins, visual-attention analysis, temporal recommendation, psychoacoustics, spatial audio, and responsible artificial intelligence. It proposes Autonomous Personalized Cinema as a computational filmmaking paradigm in which an AI system independently plans, generates, evaluates, revises, and delivers a coherent audiovisual narrative within declared human-governance constraints. A seven-layer framework is introduced, comprising a Viewer Digital Twin, Creative Intent Engine, Narrative and World Model, Multi-Agent Virtual Film Studio, Audiovisual Generation Engine, Self-Critique and Continuity System, and Personalized Exhibition Layer. The Viewer Digital Twin is extended through optional camera-based gaze mapping, behavioral observation, contextual and temporal preference modeling, and a Generative Preference Memory that stores abstract attributes of favored characters, environments, narratives, and sound experiences rather than copies of existing works. The framework rejects absolute independence from human institutions and defines autonomy specifically at the point of creative execution. It emphasizes consent, local processing, data minimization, uncertainty, viewer control, provenance, similarity screening, and protection against manipulation. The resulting framework provides a theoretical architecture, research questions, testable propositions, and a staged validation programme for the future development of autonomous AI-generated cinema.

Keywords: autonomous personalized cinema; AI-generated film; narrative review; runtime creative autonomy; viewer digital twin; visual-attention mapping; temporal preference modeling; generative preference memory; multi-agent filmmaking; psychoacoustics; responsible AI.

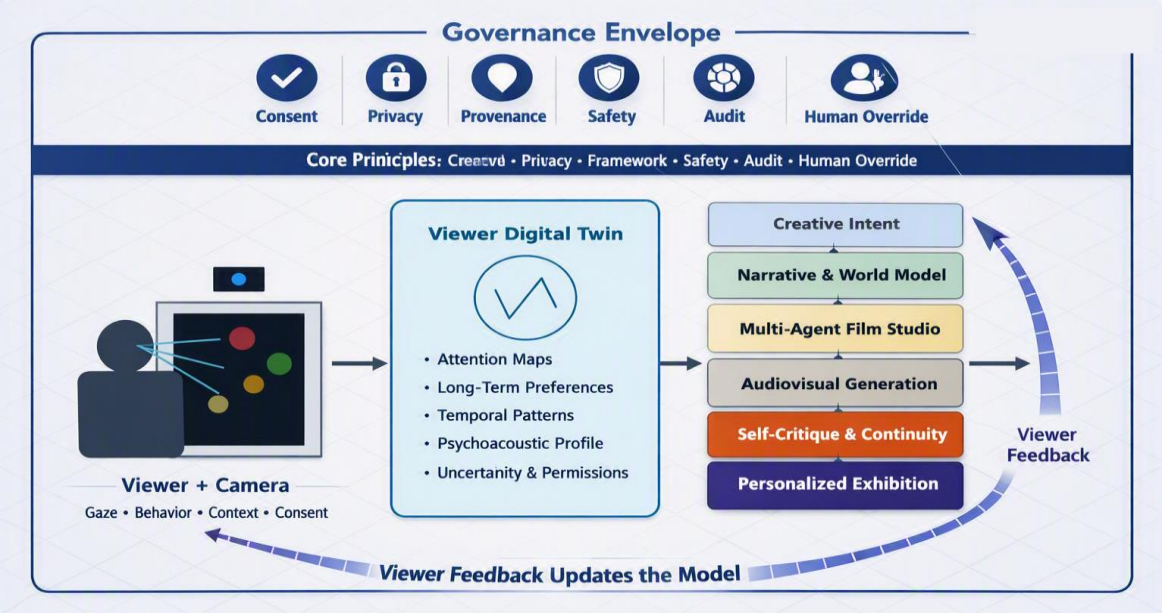


Figure 1. Graphical abstract of the Autonomous Personalized Cinema paradigm.

I. INTRODUCTION

Artificial intelligence is entering almost every stage of the cinema value chain. Current systems can support audience analysis, screenplay development, previsualization, image and video synthesis, virtual production, editing, visual effects, dubbing, music, sound design, and recommendation. Yet the presence of AI at many production stages should not be confused with an autonomous filmmaker. Most systems remain specialized tools whose goals, source materials, aesthetic direction, and acceptance criteria are supplied by people. A model that writes dialogue is not by itself a screenwriter in the operational sense; a model that generates a convincing shot is not by itself a director; and a recommender that selects an existing title has not created a film for its viewer.

Several components required for a more autonomous system are nevertheless emerging. Hierarchical story-generation systems can construct titles, characters, settings, plot structures, and dialogue, while memory-enhanced planning and critic-based approaches attempt to preserve causality over longer narratives (Mirowski et al., 2022; Bae & Kim, 2024; Wang et al., 2025). Multimodal generation models can synthesize short audiovisual sequences, edit imagery, condition outputs on user-provided references, and increasingly coordinate multiple shots (Polyak et al., 2024). Agentic production frameworks distribute tasks among simulated roles such as director, screenwriter, storyboard artist, cinematographer, and location manager (Wu et al., 2025; Xu et al., 2025). Recent benchmarks and preprints have begun to address long-horizon script faithfulness, profile-aware evaluation, and persistent entity consistency, but they also expose large remaining

failures between shots, scenes, and production stages (Chen et al., 2026; He et al., 2026; Mu et al., 2026).

A second research stream concerns personalization. Contemporary recommendation systems primarily decide which existing film to present. Personalized storytelling and profile-conditioned generation suggest a different possibility: the content itself may be generated for a particular person. MirrorStories reported higher engagement for identity- and interest-conditioned stories than for generic comparisons, while PersonaTwin and GRAVITY demonstrate richer user representations that combine demographic, behavioral, psychometric, cultural, value-based, and preference information (Yunusov et al., 2024; Chen et al., 2025; Dey et al., 2026). Recommendation-as-Generation extends this trajectory by treating generated video as the recommended item rather than limiting the system to a fixed catalogue (Cheng et al., 2026).

The conceptual step proposed in this paper is therefore larger than adaptive thumbnails, branching narratives, or the generation of isolated clips. Autonomous Personalized Cinema asks whether an AI system could construct or retrieve an authorized model of one viewer, form a creative intention, design a coherent fictional world, generate the complete audiovisual work, evaluate its own failures, revise the result, and deliver a version tailored to that viewer without continuous human creative selection during execution.

The viewer model must also move beyond a static record of genres and ratings. Viewers reveal preferences through explicit choices, but also through completion, pausing, replay, skipping, volume changes, subtitle use, session duration, and contextual patterns. A person may prefer a romantic drama on a recurring evening, an action film on a weekend, a short comedy during a work break, or a reflective film late at night. Temporal recommender research demonstrates that preferences may contain both periodic patterns and evolving patterns influenced by recent events (Cho et al., 2021). The same person may therefore require different films at different times without possessing a contradictory identity.

This study further introduces an optional camera-based visual-attention component. A camera integrated with or positioned near the viewing display may estimate approximate gaze direction and construct heat maps for faces, characters, objects, environments, lighting, and composition. Research on film viewing shows that editing, narrative comprehension, task, and scene design strongly influence gaze, while webcam-based systems can reproduce robust large-area gaze effects with lower precision than laboratory equipment (Loschky et al., 2017; Bruckert et al., 2023; Van der Cruyssen et al., 2024). Consequently, gaze cannot be interpreted as direct proof of liking. It must be

treated as uncertain evidence and combined with behavior, context, explicit feedback, and viewer correction.

A related contribution is the Generative Preference Memory. Rather than storing copies of favored actors, fictional characters, film sets, or locations, the proposed system stores abstract preference attributes. These may include apparent age range, facial structure, expression style, vocal qualities, role type, architectural form, spatial openness, light temperature, weather, camera movement, dialogue density, music intensity, and auditory comfort. The Creative Intent Engine may recombine these attributes to generate novel characters and places. This approach may reduce direct dependence on reproducing existing intellectual property, but it does not eliminate copyright, personality, trademark, privacy, digital-replica, or training-data obligations. Similarity screening, licensing, provenance, and legal review remain necessary.

The central research problem is not merely the duration of generated video. It is the preservation of creative intention, causal structure, character motivation, entity identity, performance, visual continuity, acoustic perspective, and personal relevance across a complete work, while preventing the personalization process from becoming covert psychological manipulation. Existing research addresses portions of this problem, but, within the scope of the searches described in Section II, the authors did not identify an empirically validated architecture that integrates deep viewer modeling, runtime creative intention, long-form narrative planning, multimodal production, autonomous criticism, temporal preference, visual-attention evidence, psychoacoustic adaptation, and responsible governance in one closed loop.

This narrative review therefore synthesizes the multidisciplinary knowledge required to define Autonomous Personalized Cinema and proposes a seven-layer conceptual framework. The framework is intended both as a theoretical contribution and as the architectural foundation for a future research and development programme in AI-generated cinema.

A. Aim and Contributions

The study aims to synthesize the multidisciplinary knowledge required for autonomous personalized cinema and to develop a conceptual framework and staged research programme for designing, implementing, and empirically validating an AI system capable of generating coherent films for individual viewers.

The paper makes seven principal contributions:

1. It defines runtime creative autonomy in filmmaking and distinguishes it from automation, co-creation, recommendation, interactive branching, and claims of absolute machine independence.

2. It separates personalized selection, personalized presentation, and personalized generation.

3. It proposes a seven-layer closed-loop architecture linking viewer modeling, creative intent, narrative memory, virtual production agents, audiovisual synthesis, autonomous critique, and exhibition.

4. It extends the Viewer Digital Twin through camera-based visual-attention mapping, behavioral signal fusion, temporal preference modeling, and explicit uncertainty and permissions.

5. It introduces Generative Preference Memory as an abstract, rights-aware representation of favored character, environment, narrative, and acoustic attributes.

6. It formulates research questions, testable propositions, work packages, and a staged empirical validation programme.

7. It embeds consent, privacy, provenance, viewer control, safety, non-manipulation, and legal accountability throughout the architecture rather than applying them only after generation.

II. NARRATIVE REVIEW AND FRAMEWORK-DEVELOPMENT METHODOLOGY

This study adopts a narrative literature review and conceptual framework-development approach. A narrative design was selected because autonomous personalized cinema is an emerging and multidisciplinary problem whose relevant evidence is distributed across computational creativity, narrative generation, generative video, virtual production, multi-agent systems, user modeling, recommendation, eye tracking, psychoacoustics, spatial audio, human–AI interaction, and responsible-AI governance. The purpose is not to estimate a pooled treatment effect or claim exhaustive coverage. It is to identify the capabilities required by the proposed paradigm, map demonstrated technologies to those capabilities, expose unresolved gaps, and synthesize the relationships into a research architecture.

The review combined the reference base of the initial conceptual draft with targeted searches conducted through August 2026. Sources were identified from ACL Anthology, arXiv, PubMed and

PubMed Central, SpringerLink, official conference repositories, institutional publication records, and official regulatory or standards repositories. Search expressions combined terms including AI-generated film, autonomous filmmaking, computational creativity, long-form narrative generation, multi-agent movie generation, generative video, personalized storytelling, viewer modeling, digital twin, gaze estimation, eye tracking, temporal preference, context-aware recommendation, psychoacoustics, dialogue enhancement, HRTF personalization, responsible AI, copyright, and digital replicas.

Sources were included when they contributed directly to one or more required capabilities: viewer modeling; creative-intent formation; narrative planning and memory; audiovisual generation; production-agent coordination; entity and continuity management; aesthetic or personal-fit evaluation; visual-attention measurement; temporal and contextual adaptation; psychoacoustic personalization; adaptive exhibition; or governance. Foundational works were retained where they established concepts relevant to creativity, narrative planning, user modeling, or human–AI interaction. Recent preprints were included when they represented emerging technical developments directly relevant to long-form or agentic audiovisual generation; they are explicitly treated as provisional evidence rather than settled capability.

The literature was synthesized thematically. Evidence was first organized into capability domains. Demonstrated technologies were then compared with the requirements of complete filmmaking. Recurrent deficits were classified as knowledge gaps. The seven-layer architecture emerged by grouping capabilities according to their role in the closed-loop process and by requiring persistent shared state, hierarchical control, autonomous evaluation, explicit uncertainty, and governance across the complete system.

Because the study is conceptual and narrative, it does not follow PRISMA, claim a systematic review, or report meta-analytic estimates. The resulting framework and propositions require later empirical validation.

A. Epistemic Status of Reviewed and Proposed Content

Because this review combines demonstrated evidence with conceptual extrapolation, three epistemic categories are used consistently throughout the manuscript. Demonstrated capability denotes a function that has been empirically shown in the cited literature, typically within a narrower or shorter-horizon task than complete filmmaking. Reasonable integration denotes a function that has not been directly tested but that follows plausibly from combining two or more demonstrated capabilities under stated assumptions; such claims are

technically motivated rather than empirically confirmed. Proposed capability denotes an element that is original to this framework, has not been implemented or tested in any cited source, and is presented as a target for future empirical validation. Table II marks each evidence stream according to the first category; the seven-layer framework (Section VI), the research questions and propositions (Section VII), and the validation roadmap (Section VIII) should be read throughout as reasonable integration or proposed capability rather than as findings established by the reviewed literature.

III. SCOPE AND DEFINITIONS

A. Three Levels of Personalization

The phrase personalized cinema is currently applied to systems with very different creative scope. The distinctions in Table 1 are essential because only the third level changes the film itself into a viewer-specific generated artifact.

Level	Definition	Typical examples	Central computational question
Personalized selection	An existing film is selected from a catalogue.	Recommendation, ranking, search, personalized home page.	Which available film should this viewer watch?
Personalized presentation	The underlying work remains substantially fixed, but delivery elements change.	Thumbnail, trailer, language, subtitle, dialogue enhancement, accessibility, alternate validated scenes.	How should this film be presented to this viewer?
Personalized generation	The system creates or substantially reconstructs narrative and audiovisual form for one viewer.	Viewer-specific story, characters, pacing, visual grammar, sound design, music, and ending.	What film should be created for this viewer, now, under these constraints?

Table I. Levels of cinematic personalization.

B. Operational Autonomy Rather Than Absolute Independence

A claim that an artificial filmmaker is completely independent of people would be misleading. Humans design architectures, provide or license data, establish computing access, define legal and safety boundaries, and decide whether a system may be deployed. The appropriate research target is runtime creative autonomy: independent creative execution within declared constraints and institutional accountability.

Runtime creative autonomy is defined here as the capacity of a system, after legitimate activation and without continuous human creative selection, to:

1. construct or retrieve an authorized model of the intended viewer;

2. formulate a cinematic concept, theme, genre, emotional trajectory, novelty level, and creative brief;

3. design a fictional world, characters, relationships, conflicts, causal structure, and ending;

4. translate the narrative into scenes, shots, performances, camera decisions, lighting, production design, and sound intentions;

5. generate synchronized image, motion, dialogue, ambience, effects, music, and spatial presentation;

6. edit, mix, evaluate, and revise the work against narrative, aesthetic, technical, ethical, and viewer-specific criteria; and deliver a pre-generated personal version or select among prevalidated alternatives during exhibition.

This definition separates creative execution from legal and moral responsibility. A runtime-autonomous system may make production decisions independently while remaining subject to human governance, audit, suspension, and viewer override.

C. Operationalizing Runtime Creative Autonomy: A Six-Level Matrix

The capacity list above specifies what runtime creative autonomy entails but not its degree. Because empirical comparisons will require systems that automate different proportions of the filmmaking process, autonomy is operationalized here as six cumulative levels; a system may occupy different levels for different creative decisions, for example narrative planning versus final color grading, rather than a single global level.

Level 1, Human-directed generation: the human specifies and approves every creative decision; the system executes single, human-specified instructions (e.g., a prompt-to-clip tool).

Level 2, AI-assisted generation: the human originates the creative concept and reviews outputs; the system proposes drafts or variations for human selection (e.g., co-writing tools such as Dramatron; Mirowski et al., 2022).

Level 3, AI-coordinated generation: the human sets goals and constraints and intervenes at checkpoints; specialized agents coordinate execution under a shared plan (Layer 4) without autonomous evaluation.

Level 4, Autonomous planning and generation: the human defines governance boundaries only; the system forms its own creative brief, plans, and generates without per-decision approval (Layers 1 through 5 operating without Layer 6 self-critique).

Level 5, Autonomous generation with self-evaluation and revision: the human sets objectives and audits outcomes; the system evaluates its own output against multiple criteria and revises before delivery (Layers 1 through 6). This is the level the framework targets as its primary object of empirical validation (Stages 1 through 5, Table VIII).

Level 6, Runtime adaptive autonomous generation: the human retains override and consent authority only; the system additionally adapts the delivered work during exhibition within prevalidated bounds (the complete seven-layer architecture, including Layer 7). This level is an optional, more demanding extension requiring the additional safeguards described in the governance sections below.

Distinguishing levels also clarifies RQ1: runtime creative autonomy is not a binary property of a system but a description of which decisions occur without per-instance human selection, and at which of the six levels above.

D. Working Definition

Autonomous Personalized Cinema is a computational filmmaking paradigm in which an AI system independently plans, generates, evaluates, revises, and delivers a coherent audiovisual narrative whose content and formal characteristics are adapted to an evolving, consent-based model of an individual viewer, without direct human creative intervention during runtime execution and while operating within declared governance constraints.

Several closely related terms recur throughout this paper and are used consistently as follows. Autonomous filmmaking refers broadly to any AI capacity to perform filmmaking functions without per-decision human direction, independent of whether the output is personalized. Runtime creative autonomy is the specific construct defined above and scaled by the six-level matrix in this section. Autonomous Personalized Cinema is the complete paradigm proposed in this paper, combining runtime creative autonomy with viewer-specific generation. Personalized generation is the narrower, creative-scope category defined in Table I that Autonomous Personalized Cinema instantiates; it is distinct from personalized selection and personalized presentation, which do not change the work itself. Adaptive personalization refers specifically to Layer 7's controlled, in-exhibition adjustments and should not be confused with personalized generation, which is completed before exhibition.

E. Boundaries of the Concept

- The framework does not require machine consciousness, moral agency, or human-like subjective experience.

- It does not describe a simple recommender, an interactive branching film, a prompt-to-clip generator, or a human–AI co-writing tool, although it may incorporate components from each.
- It does not require unrestricted adaptation during viewing. Real-time adaptation is limited to prevalidated alternatives and viewer-controlled sensory or pacing adjustments.
- It does not assume that a viewer profile is a complete psychological representation. The Viewer Digital Twin is task-bounded, uncertain, revisable, and permission-limited.
- It does not claim that abstract generation removes intellectual-property or personality-right obligations.

IV. THEORETICAL FOUNDATIONS

A. Computational Creativity, Intention, and Authorship

Computational creativity concerns systems whose behavior would be regarded as creative if performed by people. Boden's distinction among exploratory, combinational, and transformational creativity is useful for artificial filmmaking: a system may explore an established cinematic space, combine distant styles and narrative structures, or alter the rules of that space (Boden, 1998). Local novelty, however, is insufficient. A filmmaker must sustain an intention, make trade-offs, and revise material in relation to the whole work.

Narrative planning similarly requires more than fluent text. Riedl and Young (2010) model narrative generation as a balance between plot coherence and believable character intentionality. In cinema, the balance extends across performance, staging, editing, sound, and visual continuity. A character cannot remain psychologically coherent in the screenplay while changing face, voice, costume, emotional state, or knowledge across generated shots.

Authorship is a separate legal and philosophical problem. The framework uses operational language: the system performs functions associated with filmmaking. Whether such performance constitutes legal authorship depends on jurisdiction, the extent of human-authored expression, and provenance. Current U.S. Copyright Office guidance maintains that copyright protection requires sufficient human-authored expressive elements and that prompting alone does not automatically establish authorship (U.S. Copyright Office, 2025). Throughout this paper, statements that the system 'creates,' 'decides,' 'directs,' or 'evaluates' describe technical creative agency only; they carry no claim about legal authorship, moral

rights, or copyright ownership, which remain distinct questions addressed separately in the intellectual-property discussion below.

B. Long-Form Narrative Planning and Persistent World Memory

Long-form narrative generation remains a core bottleneck. Dramatron demonstrated hierarchical generation of titles, characters, story beats, locations, and dialogue, but was designed as a co-creative system and retained long-range semantic problems (Mirowski et al., 2022). Dynamic hierarchical outlining, memory enhancement, critic collectives, cross-lingual planning, and temporal knowledge representations improve long-form text but continue to report failures of causality, repetition, logical order, and complex story dynamics (Bae & Kim, 2024; Migal et al., 2024; Razumovskaia et al., 2024; Wang et al., 2025). Human-level narrative generation is also not established; recent analyses identify homogeneous affect, insufficient tension, and limited structural variety in machine-generated stories (Tian et al., 2024).

An autonomous film system therefore requires a persistent Narrative and World Model rather than a sequence of isolated prompts. The model must retain what has happened, what each character knows, unresolved promises, causal dependencies, object states, spatial relationships, emotional trajectories, and the conditions that govern future events. The film should be represented hierarchically as film, acts, sequences, scenes, beats, shots, and audiovisual events, with global commitments constraining local generation.

C. Multimodal Media Generation and Multi-Agent Production

Cinema coordinates language, image, motion, performance, sound perspective, music, rhythm, and spatial presentation. Current media foundation models are highly capable at local synthesis, yet reported generation contexts remain far shorter than a feature-length narrative. Movie Gen integrated high-definition video, image-conditioned personalization, editing, and synchronized audio, but its demonstrated context was measured in seconds rather than in complete films (Polyak et al., 2024).

Multi-agent architectures provide an organizational response by separating specialized responsibilities. MovieAgent assigns roles resembling director, screenwriter, storyboard artist, and location manager; Hollywood Town proposes hierarchical cross-modal orchestration; FilmAgent coordinates director, screenwriter, actor, and cinematographer functions within virtual 3D spaces (Wei et al., 2025; Wu et al., 2025; Xu et al., 2025). The labels are engineering abstractions rather than claims that agents possess subjective artistic

identities. Their value lies in modularity, persistent responsibility, traceable decisions, and the ability to return a failure to the stage where it originated.

Recent work emphasizes the gap between script and cinematic execution. Mu et al. (2026) use a script-specialized agent and continuous cross-scene generation to improve long-horizon faithfulness. EntityBench shows that character, object, and location consistency declines sharply with recurrence distance, while explicit entity memory improves fidelity (He et al., 2026). DirectorBench demonstrates the need for diagnostic evaluation across script, visual, audio, cross-modal, and stability dimensions rather than a single quality score (Chen et al., 2026).

D. Viewer Modeling, Digital Twins, and Personalized Generation

Viewing history provides useful but incomplete evidence. It reveals what was consumed, but not necessarily why a particular scene, character, environment, or sound was preferred. Personalized generation requires a richer model that may include explicit choices, behavioral patterns, culture, values, personality, tolerance for ambiguity, accessibility needs, auditory preferences, and situational context.

MirrorStories provides evidence that personalized identity and interest cues can improve engagement in short textual narratives (Yunusov et al., 2024). PersonaTwin integrates demographic, behavioral, and psychometric data into adaptive digital representations (Chen et al., 2025). GRAVITY shows that values, interests, beliefs, personality, and culture can support preference-grounded generation beyond interaction logs (Dey et al., 2026). PerSE demonstrates that open-ended outputs can be evaluated against inferred personal preferences with interpretable feedback (Wang et al., 2024). These studies do not establish personalized cinema, but they justify a task-bounded Viewer Digital Twin.

The use of digital twin terminology requires restraint. The proposed twin does not reproduce the whole person. It is a dynamic model containing only attributes required to generate and evaluate cinematic experiences. It must represent confidence, provenance, expiry, contradiction, and permission for each attribute.

E. Visual Attention During Film Viewing

Film directors use composition, faces, movement, focus, lighting, and editing to guide attention. Eye-tracking research demonstrates strong relationships between cinematic form and gaze. Bruckert et al. (2023) assembled an eye-tracking database for film sequences and reported relationships between editing and scanpaths. Loschky et al. (2017) showed that

narrative comprehension and viewing task affect gaze control, challenging the assumption that attention is determined solely by bottom-up visual features. Neurocinematic research similarly demonstrates that structured film can produce synchronized patterns across viewers, while leaving open the artistic question of whether synchronization is always desirable (Hasson et al., 2008).

Webcam-based gaze estimation offers a lower-cost route to approximate screen-level attention. WebGazer demonstrated real-time browser-based gaze estimation using ordinary webcams (Papoutsaki et al., 2016). Subsequent validations show that robust effects can be reproduced, but with noisier data and smaller effect sizes; large areas of interest are more defensible than fine-grained claims about small visual details (Van der Cruyssen et al., 2024). Some webcam systems approach laboratory performance under controlled conditions, yet calibration, head movement, screen geometry, lighting, and hardware remain important sources of error (Kaduk et al., 2024; Valtakari et al., 2024).

Attention must not be equated with liking. Gaze may indicate attraction, surprise, confusion, threat, novelty, narrative importance, or visual dominance. A gaze heat map is therefore evidence about allocation of attention, not a direct emotional label. The proposed system combines it with replay, pause, skip, completion, direct ratings, and explicit correction.

F. Temporal and Contextual Preference

Preferences are conditional. Time-aware recommender research distinguishes periodic patterns from evolving patterns caused by recent events or changing interests. TimelyRec models both forms and evaluates not only which item should be recommended but when it should be recommended (Cho et al., 2021). For cinema generation, the equivalent question is whether the system should create a particular type of film for the viewer in the present context.

The Viewer Digital Twin should therefore distinguish long-term preference, periodic preference, and session-specific intent. A recurring Thursday-evening preference for romance may coexist with a long-term preference for science fiction and a present desire for a documentary. These patterns should remain probabilistic. The system must ask for confirmation when confidence is low, and a prediction must never become a permanent label.

G. Psychoacoustic Personalization

Sound is not a downstream decoration. Cinema uses dialogue, ambience, effects, music, silence, and spatial perspective to control intelligibility, attention, anticipation, tension, fatigue, and emotional meaning. A universal mix assumes that viewers share the same hearing ability, equipment, environment, spatial perception, dynamic-range tolerance, and preference for background density. That assumption is weak.

Research demonstrates inter-individual differences in preferred gain and audio quality and shows that machine-learning methods can learn individualized settings (Jensen et al., 2019; Rajagopalan et al., 2024). Dialogue enhancement can reduce listening effort by changing the relationship between foreground speech and background sound (Torcoli et al., 2022). Individualized head-related transfer functions influence spatial localization and timbral quality in binaural reproduction (Guezenoc & Seguier, 2020; Hu et al., 2026). These findings justify a psychoacoustic profile governing dialogue balance, spectral shaping, dynamic range, sound-event intensity, music density, and spatial rendering.

Psychoacoustic adaptation must remain within artistic and safety constraints. Reducing sudden sounds for a sensitive viewer should not erase a causal event; dialogue enhancement should not destroy the intended acoustic perspective; and emotional personalization should not exploit known vulnerabilities.

Psychoacoustic personalization, as used in this paper, refers specifically to adaptations of dialogue balance, spectral shaping, dynamic range, sound-event intensity, and spatial rendering that are intended to change the cinematic experience for an individual viewer, for example by rendering a director-intended acoustic perspective differently for a viewer who reports greater sensitivity to sudden sound. It is distinct from, though it may draw on techniques from, four adjacent fields: general audio personalization, which optimizes playback for a device or environment rather than a narrative intention; accessibility-oriented audio adaptation, which restores intelligibility or comfort to a defined level for viewers with hearing loss or sensory sensitivity, typically governed by accessibility standards rather than artistic judgment; conventional sound mixing, a fixed, non-personalized creative decision applied identically to all viewers; and clinical or therapeutic audio applications, which target diagnosed hearing or auditory-processing conditions under professional supervision and fall outside the scope of this framework. A given adaptation in Layer 7 may serve either a cinematic-experience purpose, such as adjusting tension pacing through dynamic range, or

an accessibility and comfort purpose, such as raising dialogue clarity to a stated threshold; the two purposes require different validation criteria and should not be conflated in the psychoacoustic trial proposed in Table VIII.

Evidence stream	Demonstrated capability	Remaining limitation	Framework implication
Computational creativity and narrative planning	Hierarchical outlines, character planning, critic loops, memory.	Long-range causality, tension, thematic development, and payoff remain unstable.	Persistent Narrative and World Model with global commitments. (Layer 3)
Generative video and audio	High-quality local synthesis, editing, reference conditioning, synchronized sound.	Short duration, cross-shot identity drift, acoustic inconsistency.	Hierarchical audiovisual engine with persistent entity and sound state. (Layer 5)
Multi-agent production	Role specialization and production-style orchestration.	Weak shared state and uncertain error routing across stages.	Showrunner coordination, structured handoffs, and cyclic review. (Layer 4)
Viewer modeling and personalized generation	Profile-conditioned text and user digital-twin approaches.	Limited explanation of why a person prefers cinematic forms.	Multidimensional, uncertain Viewer Digital Twin. (Layer 1)
Eye tracking and film gaze	Attention can be estimated and related to editing and comprehension.	Webcam precision is limited; gaze is not equivalent to liking.	Large AOIs, confidence scores, behavioral fusion, viewer correction. (Layer 1, optional module)
Temporal recommendation	Periodic and evolving preference patterns can be learned.	Recommendation does not equal generation and context can be misread.	Temporal-context model and session confirmation. (Layer 1)
Audio personalization	Personal gain, dialogue enhancement, HRTF adaptation.	Rarely integrated into story, editing, and dramatic decisions.	Psychoacoustic profile linked to intent and final mix. (Layers 1, 5, and 7)
Responsible AI and law	Risk frameworks, consent, privacy, provenance, human oversight.	Psychologically targeted generation and digital replicas create new risks.	Governance envelope active at every layer. (governance envelope, Section IX)

Table II. Evidence streams and their implications for Autonomous Personalized Cinema.

V. Research Problem and Knowledge Gaps

The technologies required for autonomous personalized cinema are emerging across separate research communities, but, within the scope of the searches described in Section II, the authors did not identify an established and empirically validated system that integrates them into a complete filmmaking loop. The problem is not merely generating longer video. It is preserving intention and continuity across multiple representational levels while adapting the result to one viewer without reducing the work to repetitive preference confirmation or covert emotional optimization.

The central research problem is stated as follows:

Existing AI systems can automate or assist isolated stages of cinematic creation, production, evaluation, and personalization, but, to the authors' knowledge and within the scope of this review, no published architecture has been validated that unifies deep viewer modeling, visual-attention evidence, temporal preference, autonomous creative intention, long-form narrative planning, persistent multimodal continuity, complete audiovisual production, self-critique, psychoacoustic adaptation, and responsible governance into a system capable of generating a coherent film for an individual viewer.

Gap	What current systems demonstrate	What remains unresolved	Framework response
System integration	Strong modules for writing, image, video, audio, agents, editing, and recommendation.	No unified creative loop connecting viewer model, intent, narrative, production, evaluation, and revision.	Seven-layer architecture with shared state and feedback.
Long-horizon narrative	Hierarchical outlines, critics, and memory improve text.	Stable causality, motivation, thematic development, suspense, and payoff across a film.	Narrative and World Model with causal graph and commitments.
Multimodal continuity	High-quality local generation and emerging multi-shot workflows.	Identity, location, object, voice, lighting, sound perspective, and music across distant shots.	Persistent entity memory and multimodal continuity critics.
Viewer understanding	History, ratings, and profile-conditioned generation.	Why a viewer prefers specific narrative, visual, and auditory patterns.	Viewer Digital Twin with explicit, implicit, contextual, cultural, visual, and acoustic variables.
Visual preference inference	Eye tracking and gaze heat maps.	Distinguishing attraction from confusion, salience, threat, or narrative importance.	Behavioral fusion, uncertainty, and explicit correction.
Temporal preference	Time-aware recommendation and periodic behavior modeling.	Predicting what film to generate in a particular session without deterministic labeling.	Long-term, periodic, and session-specific context model.
Autonomous evaluation	Prompt alignment and local quality scores.	Judgment of dramatic value, originality, emotional arc, personal fit, and safety.	Multi-critic, human-calibrated revision system.
Psychoacoustic integration	Dialogue enhancement, gain preference, and HRTF personalization.	Integrating auditory needs into dramatic design and final mix.	Personal sound profile linked to the creative brief.
Rights-aware generation	Generative models can synthesize new-looking people and places.	Similarity to real performers, fictional characters, sets, brands, and protected assets.	Abstract preference memory, provenance, licensing, and similarity screening.

Gap	What current systems demonstrate	What remains unresolved	Framework response
Governance	General responsible-AI, copyright, privacy, and risk frameworks.	Consent and accountability for cameras, vulnerability inference, and psychologically targeted narratives.	Governance envelope, local processing, audit, limits, and viewer override.

Table III. Principal knowledge gaps and framework responses.

In summary, the reviewed literature demonstrates individually mature capabilities in narrative planning, multimodal generation, multi-agent coordination, viewer and preference modeling, visual-attention estimation, temporal recommendation, and psychoacoustic adaptation (Table II), and it identifies specific integration gaps that no single reviewed system closes simultaneously (Table III). Within the scope of the searches described in Section II, the authors did not identify a published, empirically validated architecture that integrates these capabilities into a single closed-loop, autonomous, personalized filmmaking system. The remainder of this paper is a proposed response to that gap: Sections VI and VII present an original conceptual architecture, research questions, and propositions that extend beyond, and are not established by, the evidence summarized above.

VI. PROPOSED AUTONOMOUS PERSONALIZED CINEMA FRAMEWORK

The proposed architecture treats personalized film generation as a closed-loop system rather than a linear text-to-video pipeline. Each layer creates structured outputs that constrain subsequent layers. Evaluation may return a defect to an earlier layer, and governance surrounds the complete process. The framework contains seven functional layers and a persistent shared state.



Figure 2. The seven-layer Autonomous Personalized Cinema Framework and its governance envelope.

Layer 1: Viewer Digital Twin

The Viewer Digital Twin is a dynamic, consent-based, task-bounded model of the intended viewer. It does not claim to reproduce the whole person. It contains only variables required to generate, evaluate, and deliver cinematic experiences. Every attribute should include provenance, confidence, expiry, and permission status.

The twin should preserve contradiction and context. A viewer may prefer slow cinema in one setting and rapid comedy in another. It should therefore estimate conditional preferences rather than assign permanent labels. Long-term preferences, periodic preferences, and session-specific intent must remain distinct.

The Viewer Digital Twin differs from a conventional user profile or recommender-system profile in three respects: it is generative rather than retrieval-oriented, in that its output conditions the creation of new content rather than the ranking of existing items; it is explicitly typed by uncertainty, in that each attribute carries a confidence value and provenance rather than being treated as a binary preference flag; and it is bounded to filmmaking-relevant variables rather than constituting a general-purpose behavioral profile. Its essential core comprises the explicit cinematic-preference, behavioral, narrative,

temporal-context, cultural and value, and data-permission categories in Table IV; the visual-attention and psychoacoustic categories are optional enrichments, marked accordingly in Table IV, that a deployment may omit without disabling the architecture. Core attributes should be revisited whenever new session evidence arrives; periodic and long-term estimates should be recomputed on a rolling basis, such as after each viewing session, rather than continuously, to avoid instability from a single noisy observation.

Because viewer behavior updates the twin and the twin subsequently shapes what the system generates, the architecture contains a feedback loop in which an early incorrect inference could be repeatedly reinforced rather than corrected. The framework addresses this through profile versioning, in which each update is stored as a new version rather than overwriting the prior state so that an inference can be traced and reverted; confidence-weighted aggregation, in which new evidence is combined with existing estimates in proportion to its reliability rather than replacing them outright; decay of unconfirmed inferences, in which an attribute inferred once and never corroborated loses confidence over time and is eventually excluded from the creative brief rather than persisting indefinitely; independent evidence weighting, in which gaze-derived, behavioral, and explicit-feedback estimates are reasoned about separately before fusion, consistent with RQ7 and Proposition P7, so that a single noisy channel cannot dominate the twin; and explicit viewer correction and reset, described in the governance sections below. Left unaddressed, this loop would progressively narrow the viewer's cinematic experience rather than improve its relevance, which is precisely the outcome that the controlled-novelty principle in Layer 2 and the narrative-overfitting safeguard in Table IX are designed to prevent.

Category	Examples	Evidence source	Required safeguards
Explicit cinematic preference (core)	Genre, theme, favored works, dislikes, duration, language, accessibility.	Direct settings, ratings, questionnaires.	Editable, revocable, expiry and confidence.
Behavioral viewing pattern (core)	Completion, pause, rewind, replay, skip, abandonment, volume, subtitles.	Player interaction logs.	Purpose limitation, local aggregation where possible.
Narrative profile (core)	Complexity, ambiguity, pacing, point of view, endings, humor, violence.	Explicit choice plus repeated behavior.	Avoid permanent psychological labels.
Visual-attention profile (optional)	Attention to faces, gestures, costumes, objects, architecture, landscape, lighting, composition.	Optional webcam gaze estimates and AOI maps.	Opt-in, calibration, quality thresholds, no raw video retention by default.

Category	Examples	Evidence source	Required safeguards
Explicit cinematic preference (core)	Genre, theme, favored works, dislikes, duration, language, accessibility.	Direct settings, ratings, questionnaires.	Editable, revocable, expiry and confidence.
Temporal-context profile (core)	Day, time, available duration, device, location category, individual or group viewing.	Session metadata and temporal patterns.	Probabilistic use and pre-generation confirmation.
Cultural and value profile (core)	Language, references, representation preferences, ethical boundaries.	Direct declaration and consented profile data.	No inference of sensitive traits when direct choice is sufficient.
Psychoacoustic profile (optional)	Dialogue clarity, dynamic range, sudden-sound sensitivity, spatial rendering, fatigue.	Settings, hearing-accessibility choices, optional tests.	Safety limits and no medical claims without validation.
Data permissions and uncertainty (core)	Source, age, reliability, expiry, confidence, allowed uses.	System metadata and consent ledger.	Viewer inspection, correction, reset, deletion, and export.

Table IV. Proposed Viewer Digital Twin data categories.

- ***Camera-Based Visual-Attention Mapping***

This module is an optional, experimental enrichment of the Viewer Digital Twin rather than a required component of Autonomous Personalized Cinema; a deployment may omit it entirely while retaining full Layer 1 functionality through explicit, behavioral, temporal, cultural, and psychoacoustic evidence alone. Its inclusion is justified only to the extent that it demonstrates incremental predictive value over these less intrusive signals, which is the specific comparison proposed in Proposition P7 and RQ7.

With explicit and revocable consent, a camera positioned near the display may estimate gaze direction and map attention to scene-defined areas of interest. These areas may include faces, body gestures, costumes, objects, landscapes, architecture, lighting patterns, color fields, and camera composition. The resulting representation may combine spatial heat maps with temporal fixation sequences.

Webcam systems should be used only at a level supported by their accuracy. Large screen regions and repeated patterns are more defensible than claims about small details. Calibration quality, screen size, head pose, distance, lighting, eyewear, and dropped frames should be logged. Eye-tracking reporting guidance should inform future experiments (Dunn et al., 2024).

The module must never treat a heat map as direct evidence of affection or desire. Attention can result from visual salience, surprise, threat, confusion, or narrative centrality. Preference inference should therefore combine gaze with replay, pause, skip, completion, explicit rating, and direct comparison between alternatives.

- ***Temporal and Contextual Preference Model***

The temporal model estimates when particular preferences become relevant. It may learn recurring day-of-week and time-of-day patterns, while also weighting recent interactions and current session intent. The model should output a probability distribution rather than a deterministic genre assignment. Before expensive generation begins, the viewer should be able to accept, modify, or reject the inferred brief.

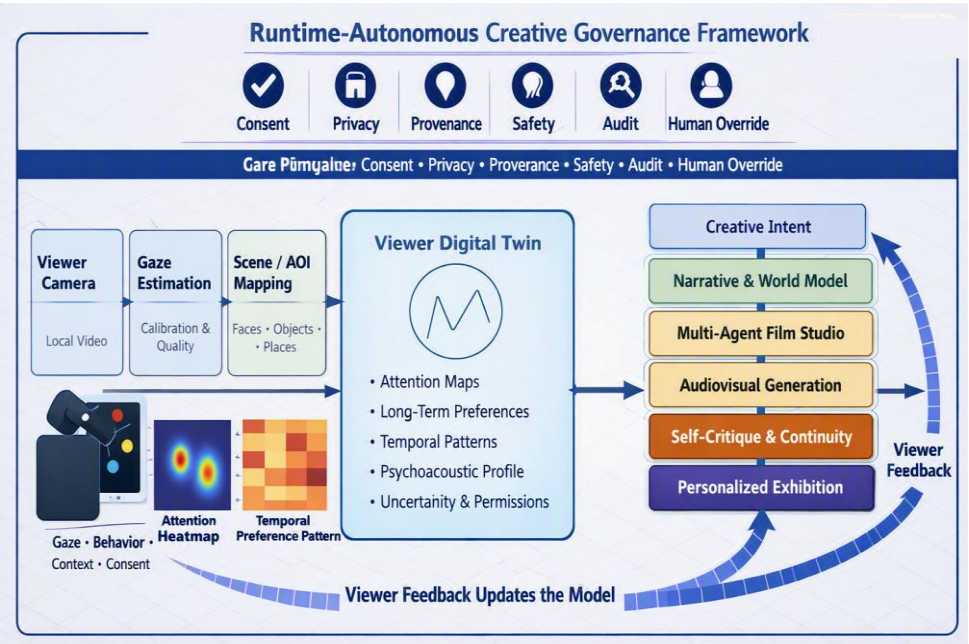


Figure 3. Integration of camera-based visual attention, behavioral evidence, temporal context, and Generative Preference Memory.

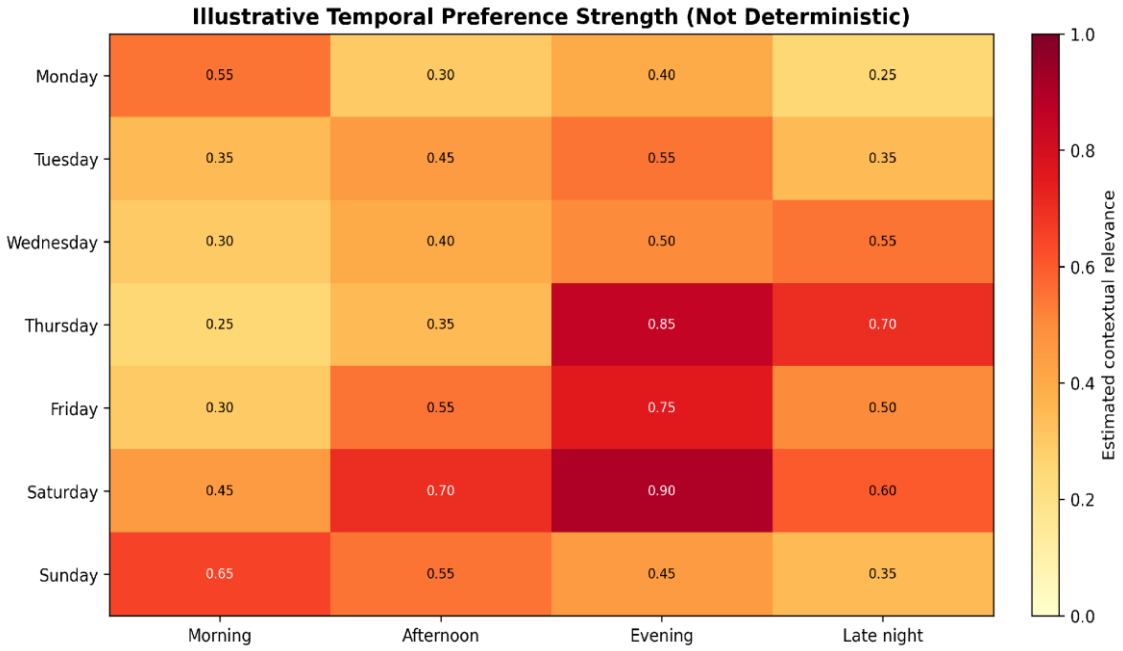


Figure 4. Illustrative weekly temporal preference map. Values are probabilistic examples, not deterministic labels.

- ***Generative Preference Memory***

The Generative Preference Memory stores abstract visual, narrative, temporal, and acoustic attributes rather than copies of previously viewed images, performers, voices, characters, or locations. A preferred character representation may contain age range, facial geometry class, expression style, movement, clothing formality, vocal range, role, and interaction pattern. An environment representation may contain architectural form, openness, historical atmosphere, weather, light direction, color temperature, texture, scale, and camera movement.

The memory must distinguish the represented entity from its cinematic presentation. Attention to a character may be caused by appearance, framing, lighting, dialogue, role, or emotional importance. These factors should remain separate where possible. The Creative Intent Engine retrieves combinations of attributes and applies controlled novelty so that the system does not reproduce the same formula repeatedly.

In implementation terms, the abstraction is intended to operate as a constrained attribute or embedding representation rather than an image-, audio-, or text-retrieval index: each stored unit is a bounded set of categorical or continuous descriptors, such as an age-range bucket, a facial-geometry class, or a color-temperature value, rather than a pointer to, or compressed encoding of, a specific viewed image, performer, or voice sample. Disentangling

these descriptors from any single source work, and applying a minimum-dissimilarity constraint between generated outputs and any individual copyrighted or identifiable reference during generation, is intended to reduce the risk of recognizable replication or memorization. Proposition P9 (Table VI) proposes to test whether this reduces recognizable similarity while preserving perceived personalization; similarity screening remains a required downstream safeguard regardless of the outcome.

Attribute-based representation does not guarantee legal independence. The generation engine must apply similarity screening for identifiable people, voices, protected fictional characters, branded designs, and recognizable production assets. Provenance should document models, licensed assets, reference materials, and human contributions.

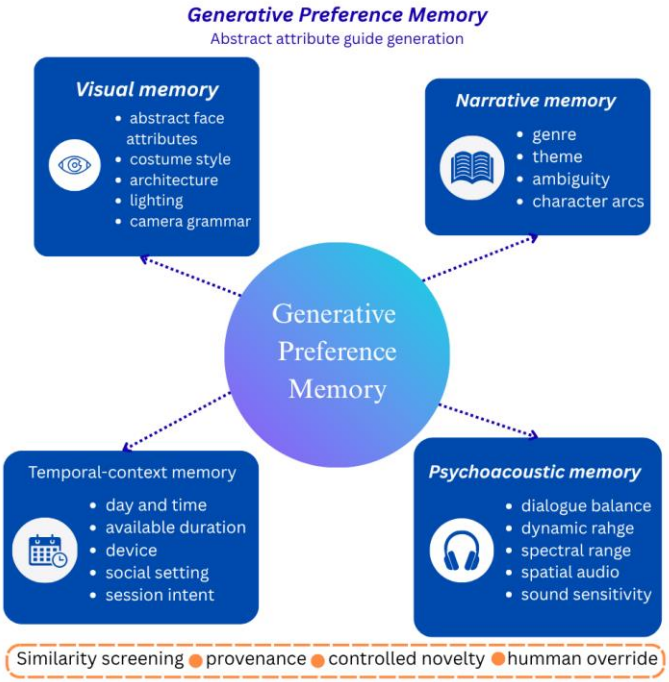


Figure 5. Generative Preference Memory as a structured store of abstract narrative, visual, temporal-contextual, and psychoacoustic attributes.

Layer 2: Creative Intent Engine

The Creative Intent Engine transforms the Viewer Digital Twin into a structured creative brief. It defines the film's thematic proposition, genre, duration, emotional trajectory, dramatic intensity, novelty, aesthetic language, sound concept, representation

requirements, and prohibited content. It also defines what must remain common across viewers when the film belongs to a shared project or franchise.

The engine should not maximize resemblance to past consumption. Controlled novelty is a central principle. Excessive personalization may produce aesthetic narrowness, predictability, and a self-confirming cultural loop. The brief should balance familiarity, exploration, surprise, and viewer-declared boundaries.

Operationally, controlled novelty can be defined as a bounded deviation between a candidate creative brief and the viewer's prior consumption pattern in the same attribute space used by the Generative Preference Memory: a novelty score below a lower bound risks the repetitive, self-confirming pattern described above, while a score above an upper bound risks abandoning the viewer's declared preferences altogether. The acceptable range is not fixed globally; it should be calibrated per viewer, since some viewers may set a wide exploration range and others a narrow one, and validated empirically at the narrative-prototype stage of the roadmap (Table VIII, Stage 2), where generic, low-novelty, and high-novelty personalized narratives can be compared on coherence, satisfaction, and surprise.

The output of the layer should be explicit and inspectable: a one-page creative constitution, a list of hard constraints, soft preferences with confidence values, a risk profile, and success criteria for later critics.

Layer 3: Narrative and World Model

This layer creates the story world and functions as persistent memory. It contains the premise, theme, characters, motivations, relationships, causal event graph, timeline, locations, objects, world rules, planted information, unresolved questions, intended payoffs, and each character's knowledge.

Before production, the layer should test whether actions are motivated, whether events are causally connected, whether the emotional arc is feasible, and whether the ending resolves the central commitments. During production, it supplies verified state information to every agent. Changes must be versioned rather than silently overwritten.

A useful implementation may combine structured knowledge graphs, timeline stores, entity cards, narrative promises, and natural-language summaries. No shot should be generated from a scene prompt alone when the persistent world state is relevant.

Layer 4: Multi-Agent Virtual Film Studio

The framework distributes responsibilities across specialized agents coordinated by a showrunner or executive-director agent. Each agent receives the approved creative constitution and verified world state, and each output is stored as a structured production artifact.

Agent	Primary responsibility	Key outputs	Examples of return conditions
Showrunner	Preserve central intention, theme, viewer contract, and global coherence.	Creative constitution, approvals, escalation decisions.	Conflicting local choices, theme drift, unsafe adaptation.
Screenwriting	Develop scene action, dialogue, subtext, and transitions.	Scene scripts, beat sheets, dialogue drafts.	Unmotivated action, exposition overload, unresolved promise.
Directing	Define performance intention, staging, shot logic, dramatic emphasis.	Blocking, performance notes, shot priorities.	Unclear action, emotional contradiction, weak scene objective.
Cinematography	Specify composition, lens logic, movement, exposure, color, and continuity.	Shot design and camera metadata.	Screen direction error, visual discontinuity, unreadable staging.
Performance	Control gesture, gaze, voice, timing, and interaction.	Performance state and voice intentions.	Identity drift, emotional mismatch, lip-sync failure.
Production design	Manage environments, costumes, props, and material world.	Entity cards, environment and prop states.	Wrong object state, costume drift, location mismatch.
Editing	Control duration, rhythm, transitions, information flow, and parallel action.	Edit decision list and pacing profile.	Causal ambiguity, poor rhythm, missing reaction, excessive duration.
Music	Develop motifs, harmonic direction, tension, release, and recurrence.	Music plan and stems.	Masking, emotional overstatement, thematic inconsistency.
Sound design	Construct ambience, effects, perspective, silence, and dialogue balance.	Sound map, stems, spatial metadata.	Acoustic geography error, intelligibility failure, unsafe level.
Continuity	Verify narrative, entities, timeline, geography, voice, and acoustics.	Continuity report and defect location.	Any cross-shot or cross-modal contradiction.
Ethics and safety	Enforce permissions, representation constraints, manipulation limits, and age rules.	Risk flags, approvals, stop decisions.	Unconsented data use, vulnerability targeting, prohibited content.
Audience simulation	Estimate interpretation and personal fit across candidate versions.	Profile-aware evaluation with uncertainty.	Low confidence, conflicting metrics, overfitting to preference.

Table V. Suggested agents in the Multi-Agent Virtual Film Studio.

Layer 5: Audiovisual Generation Engine

The generation engine translates approved plans into shots and synchronized audio. It requires persistent control of character identity, movement, spatial relationships, location appearance, object state, voice identity, language, lip synchronization, ambience, effects, music, and acoustic perspective. Generation should proceed hierarchically and commit only validated units to the film master.

Separate but linked stems should be maintained for dialogue, ambience, effects, Foley, and music. This allows personalized mixing and accessibility without regenerating the visual work. Visual assets should likewise maintain entity references, geometry, material properties, lighting state, and camera metadata.

A generation unit should not be approved solely because it is visually attractive. It must satisfy script faithfulness, entity state, cinematic intention, acoustic continuity, personal-fit constraints, rights screening, and technical quality.

Layer 6: Self-Critique and Continuity System

Creative autonomy requires the ability to reject and revise outputs. The evaluation layer should use multiple critics rather than a single aggregate score. Candidate critics include narrative coherence, character intention, originality, shot composition, performance credibility, entity continuity, audio quality, dialogue intelligibility, emotional trajectory, cultural representation, viewer alignment, and safety.

Critics must localize defects. A low score is insufficient unless the system can identify whether the failure belongs to the creative brief, scene script, blocking, generation model, edit, mix, or viewer model. DirectorBench and EntityBench support this diagnostic direction by separating dimensions and recurrence-related failures (Chen et al., 2026; He et al., 2026).

Automated evaluation must be calibrated against human judgment. Profile-aware and psychological-depth measures are more relevant than generic lexical similarity, but they remain approximations (Harel-Canada et al., 2024; Wang et al., 2024). Human experts and intended viewers are therefore required during framework validation.

Four evaluator roles should be kept distinct throughout validation. Expert evaluation, for example by screenwriters, editors, and sound designers, assesses cinematic and technical quality independently of personal relevance to any one viewer. Target-viewer evaluation assesses the specific personalized version against the individual profile it was generated for. General-viewer evaluation assesses whether a version generated for one profile remains

acceptable, coherent, and non-alienating to viewers outside that profile, which matters wherever a personalized version might later be shared or discussed. Independent safety and legal evaluation, corresponding to the ethics-and-safety agent role in Table V and supported where necessary by external human reviewers, assesses consent, similarity, manipulation, and accountability criteria on authority separate from the creative critics. Conflating these roles, for example treating a target viewer's satisfaction as evidence of cinematic quality, or treating expert approval as sufficient safety clearance, would undermine the validity of the staged programme in Table VIII.

Layer 7: Personalized Exhibition and Adaptation

Two exhibition modes are proposed. In pre-generated personalization, the complete film is rendered and validated before viewing. In controlled adaptive personalization, the system selects among prevalidated alternatives or adjusts reversible presentation variables during exhibition.

Controlled adaptation may include dialogue level, dynamic range, sensory intensity, subtitle form, music density, selected pacing options, or scene emphasis. It should not make unrestricted narrative changes that destroy causal coherence or exploit the viewer's immediate emotional state. Adaptive decisions must be logged and explainable.

The viewer should be able to select the degree of personalization, inspect which data categories are being used, pause adaptation, reset the profile, or request a common non-personalized version.

Persistent Shared State and Governance Envelope

The seven layers require a shared state containing the approved viewer contract, creative constitution, world model, entity memory, production artifacts, evaluation reports, permissions, and provenance records. Governance is not an eighth production stage. It surrounds every layer and can block an operation before data collection, generation, revision, or exhibition. Layers 1 through 6 and the persistent shared state are treated as core architectural components; within Layer 1, camera-based visual-attention mapping is an optional, experimental enrichment, and within Layer 7, controlled adaptive exhibition is an optional mode offered alongside pre-generated personalization.

VII. RESEARCH AIM, OBJECTIVES, QUESTIONS, AND PROPOSITIONS

A. Research Aim

To develop and validate a theoretical and technical architecture for runtime-autonomous, viewer-specific film generation that integrates narrative, audiovisual, behavioral, temporal, psychoacoustic, evaluative, and governance capabilities.

B. Objectives

I. Define autonomous personalized cinema and distinguish it from recommendation, interactive branching, and AI-assisted production.

II. Establish an operational definition of creative autonomy for artificial filmmaking systems.

III. Synthesize evidence across computational creativity, narrative generation, viewer modeling, digital twins, visual attention, temporal recommendation, audiovisual synthesis, psychoacoustics, and responsible AI.

IV. Develop the seven-layer framework and persistent shared-state architecture.

V. Define a consent-based visual-attention and behavioral preference module.

VI. Introduce Generative Preference Memory for abstract, controlled-novelty generation of characters and environments.

VII. Formulate research questions and propositions that support later empirical testing.

VIII. Create a staged work programme for proof-of-concept development, human evaluation, safety review, and long-form scaling.

C. Research Questions

RQ1. What distinguishes runtime creative autonomy from automation and human–AI collaboration in cinema?

RQ2. What variables must a Viewer Digital Twin contain to support coherent and ethically acceptable film personalization?

RQ3. How can preferences, culture, values, behavior, visual attention, temporal context, and psychoacoustic needs be transformed into a computational creative brief?

RQ4. What architecture can coordinate narrative planning, direction, cinematography, performance, editing, image generation, sound design, and music without direct human creative intervention during execution?

RQ5. How can long-form narrative, character, entity, visual, spatial, temporal, vocal, and acoustic continuity be maintained?

RQ6. How can the system autonomously evaluate coherence, originality, emotional impact, technical quality, safety, and viewer alignment?

RQ7. How can gaze heat maps and viewing behavior be combined to distinguish visual attention from genuine preference?

RQ8. How can periodic and session-specific context determine which type of film should be generated at a particular time without deterministic labeling?

RQ9. How can abstract preference attributes guide generation without reproducing identifiable people, voices, protected characters, or copyrighted production designs?

RQ10. Which forms of personalization should occur before exhibition, and which may be permitted during viewing?

RQ11. At what point does personalization reduce narrative integrity, aesthetic surprise, shared culture, or viewer autonomy?

RQ12. How should authorship, provenance, consent, privacy, and accountability be assigned when the complete film is generated at runtime?

D. Propositions

The following twelve propositions are theoretically motivated, testable expectations derived from the conceptual synthesis in Sections IV and VI. They are proposed hypotheses intended to guide the empirical validation programme in Section VIII; none has been tested within the present review, and none should be read as an established finding.

P1. A multidimensional Viewer Digital Twin will produce stronger perceived personalization than a profile derived solely from ratings or viewing history.

P2. Hierarchical narrative planning combined with persistent world memory will improve long-form coherence compared with direct scene-by-scene generation.

P3. A specialized multi-agent architecture will improve controllability, continuity, and error localization compared with a monolithic generation pipeline.

P4. Runtime creative autonomy depends as much on autonomous evaluation and revision as on generative capability.

P5. The relationship between personalization and cinematic quality is non-linear: insufficient personalization produces generic content, while excessive adaptation reduces coherence, novelty, and artistic identity.

P5 requires operational definitions to be testable. Personalization intensity is conceptualized as a continuum from minimal adaptation, that is, personalized selection or

presentation only (Table I), to extensive adaptation, that is, personalized generation using the fuller range of Viewer Digital Twin categories combined with controlled adaptive exhibition (Layer 7). Cinematic quality refers to expert ratings of narrative coherence, technical execution, and aesthetic control, assessed independently of personal relevance to a given viewer. Viewer satisfaction refers to the viewer's own evaluative response to the specific version received. Narrative coherence refers to causal and motivational consistency as assessed by the Layer 6 continuity critics and corroborated by expert human review. Perceived authorship or artistic value refers to the viewer's or expert's judgment of whether the work reflects a discernible, intentional creative perspective rather than an algorithmically averaged composite. Table VI operationalizes P5 as a three-level manipulation, low, moderate, and high personalization, against a common non-personalized comparison condition, measured on these four outcome constructs.

P6. Psychoacoustic personalization will improve dialogue intelligibility, listening comfort, immersion, and emotional alignment relative to a universal mix.

P7. A model combining gaze, behavioral interaction, temporal context, and explicit feedback will predict viewer preference more accurately than gaze alone.

P8. A temporally conditioned Viewer Digital Twin will produce stronger perceived relevance than a static profile that ignores periodic and session-specific preference.

P9. Generation from abstract preference attributes, combined with similarity and provenance controls, will reduce recognizable replication while preserving perceived personalization.

P10. Viewer trust will be higher when camera processing is local, raw video is not retained by default, and users can inspect, pause, correct, and delete the model.

P11. Optimization based only on completion time or immediate satisfaction will reduce diversity and may conflict with long-term viewer well-being.

P12. Governance constraints will be more effective when embedded as active architectural components than when applied only as post-generation filters.

Proposition	Independent variable or intervention	Comparison	Primary outcomes
P1	Rich Viewer Digital Twin.	History-only profile.	Perceived personalization, engagement, profile accuracy.
P2	Hierarchical plan plus persistent world memory.	Direct scene-by-scene generation.	Causal coherence, contradiction rate, expert narrative scores.

Proposition	Independent variable or intervention	Comparison	Primary outcomes
P3	Specialized agent architecture.	Monolithic pipeline.	Controllability, continuity, defect localization, revision cost.
P5	Low, moderate, and high personalization.	Common non-personalized version.	Coherence, surprise, aesthetic identity, satisfaction.
P6	Personal dialogue, dynamics, spectrum, and spatial mix.	Universal mix.	Intelligibility, listening effort, comfort, immersion.
P7	Gaze + behavior + context + explicit feedback.	Gaze only and history only.	Preference prediction, calibration, false-inference rate.
P8	Temporal-context model.	Static profile.	Session relevance, override frequency, satisfaction.
P9	Abstract attributes plus screening.	Direct reference-conditioned generation.	Recognizable similarity, novelty, personal fit, rights flags.
P10	Local processing and user controls.	Cloud processing with limited transparency.	Trust, consent acceptance, retention, willingness to use.
P12	Embedded governance agent.	Post-generation filtering.	Violation detection, prevention rate, auditability, time to intervention.

Table VI. Illustrative operationalization of selected propositions.

VIII. RESEARCH AND DEVELOPMENT ROADMAP

The framework is intended to serve as the architectural foundation for a staged Autonomous Personalized Cinema project. A defensible programme should begin with controlled short-form studies and should not claim feature-film autonomy before the component systems, preference models, continuity methods, and governance mechanisms are independently validated.

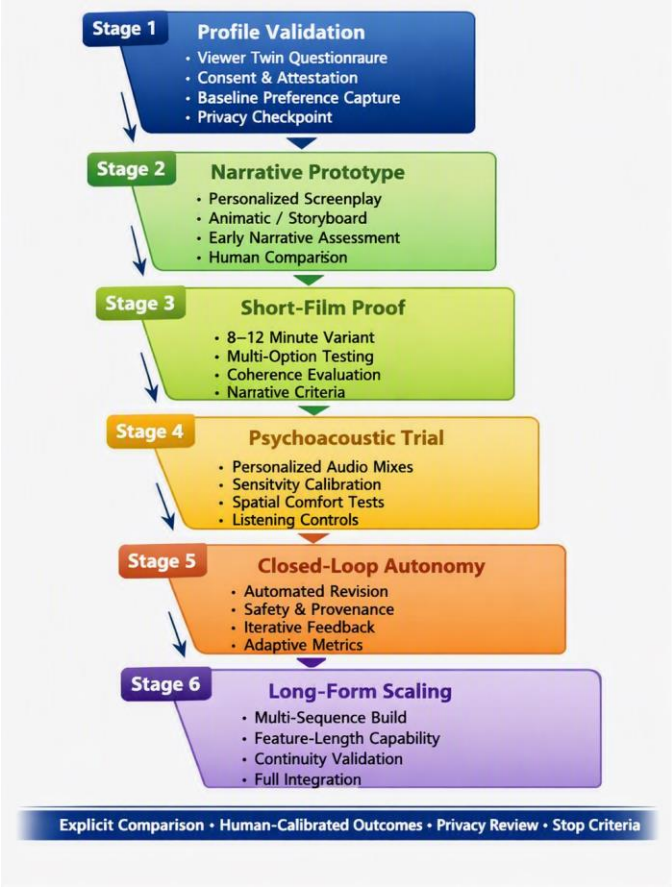


Figure 6. Staged empirical research and development programme.

Proposed Work Packages

Work package	Scope	Main deliverables	Decision gate
WP1. Viewer Digital Twin	Consent, questionnaire, behavior, gaze, temporal context, psychoacoustics.	Data schema, local attention module, profile dashboard, uncertainty model.	Can relevant preferences be measured reliably and corrected by users?
WP2. Creative Intent Engine	Translate profile and context into an inspectable film brief.	Creative constitution, controlled-novelty policy, prohibited-content rules.	Does the brief predict human judgments of desired film experience?
WP3. Narrative and World Model	Long-form planning, entities, causal graph, timeline, promises.	Narrative compiler, world-state store, contradiction detector.	Does memory improve coherence over direct generation?
WP4. Multi-Agent Film Studio	Showrunner, writing, direction, camera, performance, design, edit, sound, ethics.	Agent contracts, shared-state protocol, revision routing.	Do agents improve control and error localization?
WP5. Audiovisual Engine	Shot generation, identity, voice, audio stems, spatial and visual continuity.	Validated shot pipeline and asset memory.	Can approved scenes be generated with stable entities and synchronized audio?
WP6. Self-Critique and Continuity	Multi-critic evaluation and repair.	Diagnostic scores, revision planner, human calibration.	Do autonomous revisions improve expert and viewer ratings?

Work package	Scope	Main deliverables	Decision gate
WP1. Viewer Digital Twin	Consent, questionnaire, behavior, gaze, temporal context, psychoacoustics.	Data schema, local attention module, profile dashboard, uncertainty model.	Can relevant preferences be measured reliably and corrected by users?
WP7. Exhibition Platform	Pre-generated and controlled adaptive delivery.	Viewer controls, version comparison, adaptation log.	Can adaptation improve accessibility without harming narrative coherence?
WP8. Governance and Evaluation	Consent, privacy, provenance, rights, safety, manipulation limits.	Risk register, audit trail, similarity screening, stop procedures.	Does the system prevent rather than merely detect prohibited operations?

Table VII. Proposed project work packages.

Staged Validation Programme

Stage	Primary output	Comparison or evaluation	Purpose
1. Profile validation	Viewer Twin questionnaire, behavior model, and optional gaze module.	Prediction accuracy, test–retest stability, calibration, consent usability.	Determine whether relevant preferences can be measured reliably.
2. Narrative prototype	Personalized screenplay or illustrated animatic.	Generic versus personalized story; expert coherence and viewer engagement.	Test narrative personalization before full audiovisual cost.
3. Short-film proof of concept	8–12 minute AI-generated film in several versions.	Common, pre-personalized, and controlled-adaptive conditions.	Measure transportation, emotion, coherence, surprise, and satisfaction.
4. Psychoacoustic trial	Alternative personal mixes of the same visual film.	Universal versus personalized dialogue, dynamics, spectrum, and spatial rendering.	Test intelligibility, effort, comfort, and immersion.
5. Closed-loop autonomy	Agentic system that evaluates and revises its own film.	Human expert scores before and after automated revision.	Validate autonomous quality control and defect routing.
6. Long-form scaling	Multi-sequence or feature-length production.	Continuity decay, cost, rights audit, failure localization, safety review.	Establish whether the architecture scales beyond short films.

Table VIII Proposed staged validation programme.

Advancement criteria are specified in advance for each stage to prevent premature claims of readiness:

Stage 1, Profile validation: advance only if inferred-preference prediction exceeds a pre-registered history-only baseline, with adequate test–retest stability and consent-comprehension scores at or above a defined usability threshold.

Stage 2, Narrative prototype: advance only if expert-rated narrative coherence for personalized scripts is non-inferior to generic scripts and viewer engagement is significantly higher than the generic comparison.

Stage 3, Short-film proof of concept: advance only if the personalized and controlled-adaptive conditions do not significantly reduce coherence or safety ratings relative to the common condition, while significantly improving at least one personal-relevance outcome.

Stage 4, Psychoacoustic trial: advance only if personalized mixes improve intelligibility or comfort without reducing immersion or exceeding validated safety-level limits.

Stage 5, Closed-loop autonomy: advance only if autonomous revision improves human-expert scores relative to the pre-revision version and self-critique correctly localizes a pre-registered proportion of injected defects.

Stage 6, Long-form scaling: proceed only after an independent safety and rights review, with continuity decay and cost remaining within pre-registered bounds.

Suggested Outcome Measures

Future evaluations should combine cinematic, psychological, acoustic, behavioral, and governance measures. Candidate outcomes include narrative coherence, causal clarity, character attachment, transportation, emotional resonance, surprise, perceived personalization, aesthetic quality, willingness to continue viewing, delayed recall, dialogue intelligibility, listening effort, auditory comfort, spatial immersion, perceived manipulation, trust, consent comprehension, correction frequency, false preference inference, identifiable similarity, provenance completeness, and willingness to share personal data. Behavioral and physiological measures should be included only with explicit consent and a demonstrated need for the research question.

IX. GOVERNANCE, ETHICS, AND LEGAL ACCOUNTABILITY

Autonomous Personalized Cinema creates governance challenges that extend beyond those associated with conventional recommendation systems and AI-assisted production tools. The proposed system may construct a detailed Viewer Digital Twin, observe visual attention, identify temporal and contextual viewing patterns, generate emotionally targeted narratives, synthesize performers and voices, and modify audiovisual content without continuous human creative intervention during runtime. These capabilities may support accessibility, creativity, and viewer relevance, but they also introduce risks involving privacy, surveillance, psychological manipulation, discrimination, intellectual property, synthetic identity, cultural diversity, safety, and legal responsibility.

Governance should therefore not be implemented only as a final compliance filter after a film has been generated. It should operate as an active governance envelope surrounding the complete

architecture. Consent conditions, data permissions, prohibited uses, safety thresholds, provenance requirements, viewer rights, and accountability assignments should be represented as enforceable constraints across all seven layers of the framework.

The Creative Intent Engine should not formulate a film concept that violates the viewer's declared boundaries. The Multi-Agent Virtual Film Studio should not generate unauthorized representations or bypass performer consent. The Audiovisual Generation Engine should not use unlicensed reference assets or generate unsafe sensory outputs. The Self-Critique and Continuity System should assess ethical, legal, and psychological risks alongside narrative and technical quality. The Personalized Exhibition Layer should not intensify content or alter the film covertly merely because behavioural signals suggest that such changes may increase engagement.

This governance-by-design approach is consistent with the lifecycle orientation of the UNESCO Recommendation on the Ethics of Artificial Intelligence and the NIST AI Risk Management Framework. UNESCO emphasizes human dignity, fairness, transparency, cultural diversity, accountability, and human oversight, while NIST organizes AI risk management through the connected functions of governing, mapping, measuring, and managing risk (Tabassi, 2023; Autio et al., 2024; UNESCO, 2021).

Runtime creative autonomy should therefore be understood as the capacity to make creative decisions within an authorized institutional and ethical constitution, rather than as freedom from human accountability.

Consent, Data Minimization, and Camera Governance

The Viewer Digital Twin may contain sensitive information, including fears, values, cultural preferences, attention patterns, behavioural habits, temporal routines, accessibility requirements, and hearing-related needs. The system should collect only variables that have a defined creative, accessibility, safety, or evaluative purpose. It should explain why each data category is required, support correction and deletion, apply retention limits, and avoid inferring sensitive attributes when the same objective can be achieved through direct viewer choice.

Camera-based attention monitoring must be optional, independently controllable, and activated only through explicit opt-in consent. Consent to camera use should be separated from consent to analyse viewing history, behavioural interactions, psychoacoustic preferences, or demographic information. Refusal to activate the camera should not result in

punitive loss of access to the service. The viewer should retain access to a non-camera version and, where appropriate, to a common non-personalized version of the film. The consent interface should explain:

I. what the camera measures;

II. whether gaze direction, face position, or additional features are extracted;

III. whether any identity or emotion inference is attempted;

IV. which derived features are retained;

V. how long they are retained;

VI. whether processing occurs locally or remotely;

VII. how the information affects film generation or adaptation; and

VIII. how the viewer can pause, correct, reset, or delete the resulting profile.

Wherever technically feasible, gaze estimation, scene mapping, and feature extraction should occur locally on the viewing device. Raw camera video should be discarded immediately after the necessary features have been extracted and should not be transmitted or retained by default. Stored information should preferably consist of limited event-level features, such as approximate areas of visual attention, fixation duration, confidence values, scene identifiers, and temporal patterns, rather than complete facial recordings.

Each retained viewer attribute should have an auditable record containing its source, intended purpose, permission status, date of collection, confidence level, retention period, expiry condition, and whether it was directly stated or inferred. The viewer should be able to inspect and correct these attributes.

The system should not infer race, ethnicity, religion, political opinion, sexual orientation, health condition, criminality, personality disorder, or emotional vulnerability merely because a camera stream is available. The EU AI Act includes eye movement, facial features, voice, and other behavioural or physiological characteristics within its treatment of biometric and related AI systems and defines emotion-recognition systems as systems that infer emotions or intentions from biometric data (European Union, 2024).

The framework therefore treats gaze as probabilistic evidence of attention, not as direct evidence of preference, emotion, identity, or psychological state. A viewer may look at an object, person, or scene because it is attractive, frightening, confusing, visually dominant, narratively important, or unexpected. Gaze information should consequently be combined

with behavioural signals, explicit feedback, and contextual evidence before it is used to update the Viewer Digital Twin.

Personalization Versus Manipulation

A system capable of identifying individual preferences and generating targeted emotional experiences creates a risk that personalization may become behavioural manipulation. Conventional cinema intentionally influences attention and emotion; however, the ethical concern becomes greater when an AI system covertly exploits a particular viewer's inferred vulnerabilities or adjusts content to impair reflective choice.

Maximizing watch time, emotional arousal, completion rate, repeated use, or immediate satisfaction is not an adequate creative objective. Optimization based only on these indicators could encourage the system to intensify fear, suspense, conflict, grief, sexual content, or emotional dependency even when such escalation conflicts with viewer well-being.

The system should instead optimize a plural set of objectives, including:

A. cinematic and narrative quality;
B. viewer-declared goals;
C. meaningful novelty and surprise;
D. emotional appropriateness;
E. accessibility;
F. safety and well-being;
G. cultural diversity;
H. representational fairness;
I. long-term satisfaction; and
J. the viewer's freedom to stop or refuse adaptation.

NIST risk-management guidance emphasizes mapping, measuring, managing, and governing AI risk, while UNESCO emphasizes dignity, fairness, transparency, cultural diversity, proportionality, and human oversight (Tabassi, 2023; Autio et al., 2024; UNESCO, 2021). These principles should be translated into active architectural constraints rather than remaining general policy statements.

For example, the system should be prohibited from:

- escalating fear because anxiety is predicted to increase completion;

- generating grief-related content from inferred bereavement without explicit permission;
- exploiting loneliness to strengthen attachment to a synthetic performer;
- adapting ideological or political messages to psychological vulnerabilities;
- concealing sponsored persuasion within a personalized narrative;
- intentionally creating sensory discomfort to collect additional behavioural data; or
- preventing disengagement through emotionally coercive narrative techniques.

Controlled adaptive exhibition should preferably select among prevalidated alternatives rather than generating unrestricted emotional interventions during viewing. Significant adaptations should be logged, explainable, and reversible.

Children and viewers who may be vulnerable because of age, disability, illness, sensory sensitivity, distress, or limited digital literacy require stronger protections. Their Viewer Digital Twins should be narrower, use restrictive defaults, retain less data, and exclude attributes that are not essential to the declared creative or accessibility purpose. Camera-derived information from children should not be used for targeted advertising, psychological persuasion, or synthetic emotional dependency.

Intellectual Property, Digital Replicas, and Similarity

Runtime-generated films complicate authorship because creative contribution is distributed across system designers, model developers, training and reference data, licensed performers, Viewer Digital Twin information, autonomous agents, generated outputs, and human governance decisions. Detailed provenance is therefore necessary rather than an assumption that the AI system itself is a legal author.

Every generated film should have a traceable production record that identifies:

a. The responsible legal or institutional entity;
b. The models and model versions used;
c. Licensed datasets and reference materials;
d. Prompts and structured creative briefs;
e. Human-authored contributions;
f. Agent decisions and revision histories;
g. Performer, voice, music, and asset licenses;
h. Similarity-screening results;

i. Personalization attributes used;

j. Adaptive exhibition decisions;

k. The final release authorization.

Digital performers and synthetic voices create legal and ethical issues that may involve identity, consent, privacy, publicity, personality, contractual, labour, consumer-protection, and digital-replica rights. The public availability of an image or voice recording should not be interpreted as permission to generate a synthetic performance.

Consent for a digital performer should define the permitted performance, project scope, duration, territory, modification rights, model-training rights, transferability, compensation, prohibited uses, withdrawal conditions, and posthumous use where relevant.

The U.S. Copyright Office has recommended stronger protection against unauthorized digital replicas and has separately maintained that copyrightability depends on the presence of sufficient human-authored expressive elements. Prompting alone does not automatically establish authorship, and complex AI-generated output should not be assumed to qualify for protection merely because it required substantial computation (U.S. Copyright Office, 2024, 2025).

Generative Preference Memory may reduce direct copying by storing abstract attributes rather than complete images, voices, characters, or locations. For example, the memory may store preferences for a particular age range, restrained facial expression, warm architectural lighting, narrow historic streets, slow camera movement, or a low-pitched voice. However, recognizable similarity may still emerge from combinations of features.

A generated face may resemble a real actor. A generated voice may reproduce a distinctive vocal identity. A combination of costume, movement, name, visual design, and narrative role may reproduce a protected fictional character. A generated environment may replicate the distinctive expression of an existing production design.

The system should therefore screen generated content for similarity involving:

- human faces and bodies;
- recognizable voices;
- protected fictional characters;
- costumes and production designs;

- logos and trademarks;
- music and recurring themes;
- distinctive dialogue;
- cinematic compositions;
- branded objects; and
- restricted or licensed locations.

Similarity screening should not be treated as automatic legal clearance. Ambiguous cases should be routed to qualified human review. The applicable threshold may vary according to jurisdiction, commercial purpose, degree of similarity, public interest, and the type of right involved.

Autonomous Personalized Cinema should not assume that abstract generation removes intellectual-property obligations. Compliance requires licensed data and references, provenance documentation, similarity controls, performer consent, and jurisdiction-specific legal assessment.

Viewer Control and Shared Culture

A different film for every viewer may improve relevance, accessibility, and emotional alignment, but it may also weaken the shared cultural object that supports collective discussion, criticism, preservation, age classification, festival selection, education, and historical memory.

In conventional cinema, multiple viewers encounter substantially the same work. In Autonomous Personalized Cinema, each person may encounter a different sequence, ending, performer, environment, sound mix, or emotional emphasis. This creates uncertainty about what constitutes the film as a stable cultural object.

The platform should therefore support three clearly distinguishable modes:

a) **Common authored version:** a stable film shared by all viewers.
b) **Transparent personalized version:** a film generated or modified using declared preference categories.
c) **Controlled adaptive version:** a film that selects among prevalidated alternatives during exhibition.

The viewer should be able to:

- select the preferred mode;
- access a non-personalized version;
- disable the camera;
- pause adaptation;
- inspect the categories of information being used;
- review major changes made to the film;
- compare personalized and common versions;
- correct inferred preferences;
- exclude specific data categories;
- set the degree of personalization;
- reset the Viewer Digital Twin;
- export or delete profile data; and
- request an explanation of significant adaptive decisions.

The right to refuse personalization is a central design requirement, not a peripheral setting. A viewer who rejects camera use, behavioural profiling, or personalized generation should not be denied reasonable access to the underlying service.

The system should also avoid narrative overfitting. Repeatedly generating only what the viewer has preferred in the past may reduce surprise, originality, cultural exposure, and artistic identity. The Creative Intent Engine should therefore include controlled novelty and diversity objectives, allowing the system to remain relevant without trapping the viewer inside a repetitive aesthetic or ideological loop.

1010 **Governance Principles**

Risk	Example	Architectural control	Viewer or institutional right
Unconsented surveillance	Camera or behavioural capture is activated by default or without meaningful notice.	Explicit opt-in, visible camera status, local processing, granular permissions, and no raw-video retention by default.	Disable the camera and access a common or non-camera version.
Excessive data collection	The Viewer Digital Twin stores information unrelated to film generation or evaluation.	Purpose limitation, data minimization, retention schedules, expiry rules, and feature-level storage.	Inspect, correct, export, and delete stored attributes.
False preference inference	Threat, surprise, confusion, or narrative importance is interpreted as liking.	Multimodal evidence fusion, confidence scores, contradiction storage, explicit confirmation, and correction mechanisms.	Inspect and edit inferred preferences.
Manipulation	Fear, grief, loneliness, or vulnerability is used to maximize completion or dependence.	Prohibited optimization objectives, well-being constraints, ethics-and-safety agent, adaptive limits, and audit logs.	Request an explanation, stop adaptation, complain, and delete related data.
Biometric or sensitive inference	Identity, political views, health, emotion, or psychological vulnerability is inferred from face or gaze.	Purpose limitation, restricted feature extraction, no emotion recognition by default, and separation of gaze from sensitive inference.	Refuse the category and request deletion of derived data.
Child or vulnerable-viewer exploitation	A synthetic character pressures a child or distressed viewer into disclosure or emotional dependency.	Restricted profiles, age-appropriate design, stronger defaults, limited retention, and independent oversight.	Adult or institutional review, withdrawal, and redress.
Digital replica	A generated performer or voice resembles a real person without valid permission.	Licensed references, performer contracts, similarity screening, provenance, and human review.	Notice, challenge, removal, and compensation where applicable.
Copyright or trademark conflict	A protected character, set, costume, logo, musical theme, or branded object is reproduced.	Asset and model licensing, provenance records, similarity screening, trademark filters, and legal escalation.	Rights-holder review, complaint, and takedown process.
Narrative overfitting	Every film repeats previous preferences and eliminates novelty or cultural exposure.	Controlled novelty, diversity objectives, uncertainty preservation, and common-version availability.	Set the level and categories of personalization.
Unsafe sound	A personalized mix creates excessive level, discomfort, masking, or misleading accessibility claims.	Validated level limits, psychoacoustic testing, reversible controls, and access to the approved master mix.	Reset immediately to the validated master version.
Opaque adaptation	The film changes during viewing without disclosure or reconstructable records.	Adaptation logs, prevalidated alternatives, visible personalization status, and explanation mechanisms.	Pause adaptation and compare versions.
Cultural or representational bias	Viewer history reinforces stereotypes or excludes cultural and demographic groups.	Diverse evaluation datasets, representation review, fairness testing, and controlled cultural novelty.	Report harmful representation and request review or regeneration.
Accountability gap	No organization accepts responsibility because decisions were made by autonomous agents.	Named deployer, release authority, audit trail, model and data provenance, suspension authority, and incident-response procedure.	Independent review, complaint, remedy, and evidence preservation.
Security and profile misuse	Viewer preferences, facial features, or generated films are stolen, modified, or reused without authorization.	Encryption, access control, data separation, least privilege, monitoring, and incident-response plans.	Notification, deletion, correction, and institutional redress.

1011 *Table IX Governance risks and proposed architectural controls*

The ethical legitimacy of Autonomous Personalized Cinema depends not on eliminating creative autonomy, but on defining where autonomy may operate and where human responsibility must remain non-delegable. The system may autonomously generate, evaluate, and revise a film, but it should not autonomously redefine consent, expand its own purpose, infer prohibited sensitive attributes, exploit a vulnerability, authorize an unlicensed digital replica, conceal adaptation, or erase its production history.

The governance envelope therefore establishes four non-delegable conditions:

1. the viewer retains authority over personal data and personalization;
2. identifiable individuals retain rights over their synthetic identities and performances;
3. a named institution remains accountable for deployment, release, complaint handling, and remedy; and
4. significant creative, adaptive, data-related, and governance decisions remain traceable and open to audit.

These conditions do not weaken runtime creative autonomy. They define the legitimate technical, legal, and social environment within which such autonomy can be scientifically evaluated and responsibly deployed.

Governance Decision Hierarchy

The governance envelope is not a single filter but a layered decision structure comprising six elements. Hard constraints are non-negotiable and the system cannot override them under any optimization pressure; they include the prohibited practices listed above, the restriction on inferring sensitive attributes from camera or behavioral data, and the requirement for explicit consent before camera activation. Soft constraints are optimized within bounds; controlled novelty, dramatic intensity, and pacing, for example, may be adjusted to satisfy viewer preference and creative-quality objectives simultaneously. Escalation conditions route a decision to human review rather than resolving it autonomously; these include low-confidence similarity-screening results and any conflict between two governance rules that the system cannot resolve using its declared priority order.

A declared priority order resolves conflicts among competing rules: viewer safety and non-manipulation constraints take precedence over personalization objectives, which take precedence over novelty and efficiency objectives, and a rule lower in this order may never be satisfied by violating a rule higher in it. Stop conditions halt generation or exhibition

entirely, for example on failure to verify consent status or detection of a probable digital-replica violation, and require human authorization to resume. Final decision authority for release, suspension, and remedy rests with the named accountable institution identified in the production record, never with an autonomous agent. This hierarchy operationalizes the ethics-and-safety agent role in Table V and gives the Layer 6 critics an explicit basis for routing a flagged output to revision, escalation, or rejection rather than a single pass/fail score.

X. DISCUSSION

The framework changes the debate from whether AI possesses human-like imagination to whether a computational system can satisfy the operational requirements of filmmaking. It does not require the system to feel grief in order to stage grief. It does require the system to maintain a defensible dramatic intention, produce coherent performances, control audiovisual form, anticipate likely interpretation, and revise failures. Whether this operational achievement should be called authorship remains a legal and philosophical question; whether it can be measured as filmmaking is an empirical question.

The strongest evidence for feasibility is the convergence of previously separate research directions. Hierarchical narrative generation supplies planning and memory. Media foundation models supply image and sound synthesis. Multi-agent systems supply modular production roles and feedback. Digital twins and profile-grounded generation supply richer individual models. Eye tracking supplies approximate evidence about allocation of attention. Temporal recommender research supplies periodic and evolving context. Personalized evaluation supplies criteria that vary by viewer. Audio-personalization research supplies methods for adapting dialogue, spectrum, spatial rendering, and comfort. The unresolved contribution is integration.

Integration introduces new failure propagation. An inaccurate gaze estimate may become a false preference; the false preference may influence the creative brief; the brief may generate a recurring character type; and the evaluation agent may incorrectly reward the result because it relies on the same contaminated profile. The framework must therefore preserve uncertainty, retain intermediate representations, separate evidence sources, and make every consequential decision auditable.

The Generative Preference Memory is a particularly important boundary object. It translates viewer history into abstract creative attributes, but it must avoid both memorization and aesthetic stagnation. The same mechanism that protects against direct copying may also flatten culture into

measurable features. Controlled novelty, common versions, independent artistic constraints, and human evaluation are therefore necessary to preserve surprise and shared meaning.

The camera-based component should be treated as a research instrument, not as a premise that every screen should monitor every viewer. Its scientific value depends on whether attention data improves preference inference beyond behavior and explicit feedback. Its social acceptability depends on local processing, transparency, limited retention, and the right to refuse. P10 is therefore not a secondary usability proposition; it is a condition for legitimate deployment.

The proposed architecture also changes the industrial meaning of a film. Conventional cinema distributes one master work to many viewers. Autonomous personalized cinema may generate a family of related works or a unique artifact for each person. This affects production economics, preservation, criticism, festival selection, censorship, ratings, contracts, performer rights, and cultural memory. These effects justify treating the field as more than an extension of recommendation systems.

For the ENOSH research programme, the framework provides a sequence of testable components rather than an all-or-nothing claim. The first defensible project is not a fully autonomous feature. It is a controlled short film with a transparent Viewer Digital Twin, a persistent narrative model, a small agent team, separated audio stems, human-calibrated critics, and explicit governance. Each later stage should proceed only when the previous stage meets its validation and safety gate.

XI. LIMITATIONS

This paper is theoretical and does not demonstrate a complete autonomous film system. Several of the most directly relevant long-video and agentic-production studies are recent preprints and should be treated as emerging evidence rather than settled industrial capability. The framework combines findings from text generation, short-form video, virtual production, recommendation, eye tracking, hearing technology, and spatial audio; transfer from these domains to complete cinema must be tested rather than assumed.

The narrative review is transparent but not systematic or exhaustive. It does not estimate publication bias or pooled effects. Future studies should conduct focused systematic reviews for specific components such as webcam gaze validity, long-form continuity, personalized evaluation, and psychoacoustic adaptation.

The Viewer Digital Twin risks oversimplifying identity. Preferences are context-dependent, may change, and may be intentionally inconsistent. Gaze data are noisy and attention is not equivalent to

preference. The framework therefore requires uncertainty and direct correction, but the adequacy of those controls remains empirical.

Finally, the term autonomy is restricted to runtime creative execution. It does not imply consciousness, moral agency, or freedom from human-designed infrastructure and constraints.

XII. CONCLUSION

Autonomous personalized cinema is not achieved by extending the duration of generated video. It requires an integrated system capable of modeling a viewer, forming a creative intention, constructing a persistent narrative world, coordinating specialized production agents, generating synchronized image and sound, evaluating its own work, revising defects, and delivering a transparent personal experience within strict governance boundaries.

The framework extends personalization beyond history and ratings through optional visual-attention evidence, behavioral signal fusion, temporal context, and a Generative Preference Memory of abstract character, environment, narrative, and acoustic attributes. These additions make the system more capable, but also more ethically demanding. Camera data, inferred preferences, emotional targeting, and synthetic identities require explicit consent, local processing, uncertainty, provenance, similarity screening, viewer override, and accountable human institutions.

The individual technologies required for this transition are emerging but remain fragmented. The seven-layer framework, research questions, propositions, work packages, and staged programme proposed here provide a theoretical and practical foundation for investigating whether runtime-autonomous, viewer-specific, psychoacoustically informed, and ethically governed filmmaking can be built and validated.

DECLARATIONS

Funding: No external funding has been declared for this conceptual study.

Conflict of Interest: The author declares no personal financial conflict of interest. Because the author is affiliated with the journal's founding institution, any submission to ENOSH Science Journal should be managed by an independent editor, with the author excluded from peer-review assignment and the final editorial decision.

Data Availability: Not applicable. This narrative review and conceptual framework do not report an original dataset.

Ethics Approval: Not applicable to the present conceptual study. Future camera-based, behavioral, or psychoacoustic experiments will require prior ethics approval, informed consent, data-protection assessment, and participant withdrawal procedures.

Author Statement on AI-Assisted Drafting: Generative artificial-intelligence tools were used to assist with language refinement, structural organization, figure drafting, and literature discovery. The author is responsible for verifying the scientific claims, references, interpretations, and final manuscript and assumes full responsibility for the published content.

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human–AI interaction. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Paper 3. https://doi.org/10.1145/3290605.3300233

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., & Roberts, K. (2024). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Bae, M., & Kim, H. (2024). Collective critics for creative story generation. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 18784–18819. https://doi.org/10.18653/v1/2024.emnlp-main.1046

Boden, M. A. (1998). Creativity and artificial intelligence. Artificial Intelligence, 103(1–2), 347–356. https://doi.org/10.1016/S0004-3702(98)00055-1

Bruckert, A., Christie, M., & Le Meur, O. (2023). Where to look at the movies: Analyzing visual attention to understand movie editing. Behavior Research Methods, 55, 2940–2959. https://doi.org/10.3758/s13428-022-01949-7

Chen, J. (2007). Flow in games (and everything else). Communications of the ACM, 50(4), 31–34. https://doi.org/10.1145/1232743.1232769

Chen, J., Chen, Q., Zhang, J., Wu, Y., Li, Y., Zhang, X., Zhou, W., & Ma, C. (2026). DirectorBench: Diagnosing long-form video generation with personalized multi-agent evaluation. arXiv. https://doi.org/10.48550/arXiv.2605.30090

Chen, S., Lalor, J. P., Yang, Y., & Abbasi, A. (2025). PersonaTwin: A multi-tier prompt conditioning framework for generating and evaluating personalized digital twins. Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²), 774–788. https://aclanthology.org/2025.gem-1.66/

Cheng, Y., Wang, B., Zhang, H., Gao, X., Yin, Z., Xue, B., Li, Y., Xue, J., Ma, Y., Wang, M., Li, J., Xu, T., Liu, Z., Lin, X., Wen, S., Li, C., Liu, L., Chen, Q., Jiang, P., & Gai, K. (2026).

Recommendation as generation: Unifying personalized video generation and recommendation at industrial scale. arXiv. https://doi.org/10.48550/arXiv.2606.25496

Cho, J., Hyun, D., Kang, S., & Yu, H. (2021). Learning heterogeneous temporal patterns of user preference for timely recommendation. Proceedings of the Web Conference 2021, 1274–1283. https://doi.org/10.1145/3442381.3449947

Dey, P., Rosa, D., Zheng, W., Barcklow, D., Zhao, J., & Ferrara, E. (2026). GRAVITY: A framework for personalized text generation via profile-grounded synthetic preferences. Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, 7416–7436. https://doi.org/10.18653/v1/2026.eacl-long.348

Dunn, M. J., Alexander, R. G., Amiebenomo, O. M., Arblaster, G., Atan, D., Erichsen, J. T., Ettinger, U., et al. (2024). Minimal reporting guideline for research involving eye tracking (2023 edition). Behavior Research Methods, 56, 4351–4357. https://doi.org/10.3758/s13428-023-02187-1

European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union.

European Union. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union.

Guezenoc, C., & Seguier, R. (2020). HRTF individualization: A survey. arXiv. https://doi.org/10.48550/arXiv.2003.06183

Harel-Canada, F. Y., Zhou, H., Muppalla, S., Yildiz, Z. S., Kim, M., Sahai, A., & Peng, N. (2024). Measuring psychological depth in language models. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 17162–17196. https://doi.org/10.18653/v1/2024.emnlp-main.953

Hasson, U., Landesman, O., Knappmeyer, B., Vallines, I., Rubin, N., & Heeger, D. J. (2008). Neurocinematics: The neuroscience of film. Projections, 2(1), 1–26. https://doi.org/10.3167/proj.2008.020102

He, R., Wei, M., Yang, Z., & Ordonez, V. (2026). EntityBench: Towards entity-consistent long-range multi-shot video generation. arXiv. https://doi.org/10.48550/arXiv.2605.15199

Hu, D., Hu, J., & Jiang, C. (2026). Graph neural field with spatial-correlation augmentation for HRTF personalization. Proceedings of the AAAI Conference on Artificial Intelligence, 40(21), 17490–17498. https://doi.org/10.1609/aaai.v40i21.38803

Jensen, N. S., Hau, O., Nielsen, J. B. B., Nielsen, T. B., & Legarth, S. V. (2019). Perceptual effects of adjusting hearing-aid gain by means of a machine-learning approach based on individual user preference. Trends in Hearing, 23. https://doi.org/10.1177/2331216519847413

Kaduk, T., Goeke, C., Finger, H., & König, P. (2024). Webcam eye tracking close to laboratory standards: Comparing a new webcam-based system and the EyeLink 1000. Behavior Research Methods, 56, 5002–5022. https://doi.org/10.3758/s13428-023-02237-8

Loschky, L. C., Larson, A. M., Magliano, J. P., & Smith, T. J. (2017). What is the role of the film viewer? The effects of narrative comprehension and viewing task on gaze control in film. Cognitive Research: Principles and Implications, 2, 46. https://doi.org/10.1186/s41235-017-0080-5

Migal, A., Seredina, D., Telnina, L., Nazarov, N., Kolmogorova, A., & Mikhaylovskiy, N. (2024). Overview of Long Story Generation Challenge (LSGC) at INLG 2024. Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges, 47–53. https://doi.org/10.18653/v1/2024.inlg-genchal.4

Mirowski, P., Mathewson, K. W., Pittman, J., & Evans, R. (2022). Co-writing screenplays and theatre scripts with language models: An evaluation by industry professionals. arXiv. https://doi.org/10.48550/arXiv.2209.14958

Mu, C., He, X., Yang, Q., Chen, W., Yao, J., Liu, H., Yi, Z., Zhao, B., Chen, X., Ma, R., Ye, F., Yang, E., Deng, C., Tu, Z., Li, X., & Linus. (2026). The script is all you need: An agentic framework for long-horizon dialogue-to-cinematic video generation. arXiv. https://doi.org/10.48550/arXiv.2601.17737

Papoutsaki, A., Sangkloy, P., Laskey, J., Daskalova, N., Huang, J., & Hays, J. (2016). WebGazer: Scalable webcam eye tracking using user interactions. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 3839–3845.

Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., et al. (2024). Movie Gen: A cast of media foundation models. arXiv. https://doi.org/10.48550/arXiv.2410.13720

Rajagopalan, R., Wei, Y.-L., & Choudhury, R. R. (2024). Sample-constrained black box optimization for audio personalization. Proceedings of the AAAI Conference on Artificial Intelligence, 38(9), 10120–10128. https://doi.org/10.1609/aaai.v38i9.28881

Razumovskaia, E., Maynez, J., Louis, A., Lapata, M., & Narayan, S. (2024). Little Red Riding Hood goes around the globe: Crosslingual story planning and generation with large language

models. Proceedings of LREC-COLING 2024, 10616–10631. https://aclanthology.org/2024.lrec-main.929/

Riedl, M. O., & Young, R. M. (2010). Narrative planning: Balancing plot and character. Journal of Artificial Intelligence Research, 39, 217–268. https://doi.org/10.1613/jair.2989

Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1

Tian, Y., Huang, T., Liu, M., Jiang, D., Spangher, A., Chen, M., May, J., & Peng, N. (2024). Are large language models capable of generating human-level narratives? Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 17659–17681. https://doi.org/10.18653/v1/2024.emnlp-main.978

Torcoli, M., Robotham, T., & Habets, E. A. P. (2022). Dialogue enhancement and listening effort in broadcast audio: A multimodal evaluation. arXiv. https://doi.org/10.48550/arXiv.2207.14240

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization.

U.S. Copyright Office. (2024). Copyright and artificial intelligence, Part 1: Digital replicas. Library of Congress.

U.S. Copyright Office. (2025). Copyright and artificial intelligence, Part 2: Copyrightability. Library of Congress.

Valtakari, N. V., Hessels, R. S., Niehorster, D. C., Viktorsson, C., Nyström, P., Falck-Ytter, T., Kemner, C., & Hooge, I. T. C. (2024). A field test of computer-vision-based gaze estimation in psychology. Behavior Research Methods, 56, 1900–1915. https://doi.org/10.3758/s13428-023-02125-1

Van der Cruyssen, I., Ben-Shakhar, G., Pertzov, Y., Guy, N., Cabooter, Q., Gunschera, L. J., & Verschuere, B. (2024). The validation of online webcam-based eye-tracking: The replication of the cascade effect, the novelty preference, and the visual world paradigm. Behavior Research Methods, 56, 4836–4849. https://doi.org/10.3758/s13428-023-02221-2

Wang, D., Yang, K., Zhu, H., Yang, X., Cohen, A., Li, L., & Tian, Y. (2024). Learning personalized alignment for evaluating open-ended text generation. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 13274–13292. https://doi.org/10.18653/v1/2024.emnlp-main.737

Wang, Q., Hu, J., Li, Z., Wang, Y., Li, D., Hu, Y., & Tan, M. (2025). Generating long-form story using dynamic hierarchical outlining with memory-enhancement. Proceedings of NAACL 2025, 1352–1391. https://doi.org/10.18653/v1/2025.naacl-long.63

Wei, Z., Li, M., Zhang, Z., Yuan, R., Hui, P., Qu, H., Evans, J., Agrawala, M., & Rao, A. (2025). Hollywood Town: Long-video generation via cross-modal multi-agent orchestration. arXiv. https://doi.org/10.48550/arXiv.2510.22431

Wu, W., Zhu, Z., & Shou, M. Z. (2025). Automated movie generation via multi-agent CoT planning. arXiv. https://doi.org/10.48550/arXiv.2503.07314

Xu, Z., Wang, L., Wang, J., Li, Z., Shi, S., Yang, X., Wang, Y., Hu, B., Yu, J., & Zhang, M. (2025). FilmAgent: A multi-agent framework for end-to-end film automation in virtual 3D spaces. arXiv preprint.

Yunusov, S., Sidat, H., & Emami, A. (2024). MirrorStories: Reflecting diversity through personalized narrative generation with large language models. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 6702–6717. https://doi.org/10.18653/v1/2024.emnlp-main.382

APPENDIX A. MINIMUM DATA AND DECISION RECORD FOR A PROTOTYPE

A proof-of-concept system should record the following information for every generated film or validated branch:

- viewer consent version and enabled data categories;
- profile attributes used, confidence, source, expiry, and viewer corrections;
- camera-calibration quality and retained attention features, without raw video by default;
- creative constitution and controlled-novelty settings;
- narrative and world-model version;
- agent outputs and approvals;
- generation models, licensed assets, references, prompts, seeds, and transformations;
- entity, continuity, audio, safety, and similarity evaluation reports;
- revision history and reasons for rejection or regeneration;

- personalized mix or adaptation decisions; and
- viewer feedback, deletion requests, overrides, and model-update decisions.