Article

A Methodological Framework for Training Tool-Aware AI Entertainers for Live Streaming

Jinsen Wu [1,*]

[1] HockeyStack corporate office, San Francisco, California, 94107, United States
* Correspondence: Jinsen Wu, HockeyStack corporate office, San Francisco, California, 94107, United States

Abstract: This paper introduces a methodological framework for training AI systems to function as tool-aware entertainers in live streaming environments. Moving beyond generic language models that generate scripted responses, the framework conceptualizes tool awareness as a triadic competence: the ability to perceive available digital tools---such as real-time graphics engines, audience sentiment analyzers, and interactive overlay systems; to select among them dynamically based on both performative goals and unfolding audience conditions; and to orchestrate their use within the strict latency and coherence constraints of live, multimodal performance. Grounded in an interdisciplinary synthesis of performance theory, cognitive science, and machine learning, the proposed architecture comprises three interdependent layers: a Tool Perception Layer that embeds tool capabilities from documentation and interface signals; a Performative Intent Mapping Layer trained on annotated live-stream data to link social intentions---like building rapport or escalating excitement---to appropriate tool combinations; and a Runtime Orchestration Layer that schedules tool invocations with built-in fallback logic and timing safeguards. The framework is supported by a curated, multimodal training corpus drawn from diverse streaming genres and regions, and evaluated through a dual-axis protocol assessing both functional reliability and performative appropriateness. Implementation considerations address platform-specific constraints, cultural calibration for regional norms---particularly in East Asian streaming ecologies---and ethical integration of human oversight as a structural component of the training loop. The work does not deliver a deployed system but advances a domain-specific scaffolding for engineering AI performers whose actions are intentional, contextually grounded, and aesthetically coherent. It positions tool awareness not as technical utility but as a form of performative cognition essential to meaningful human--AI co-presence in real-time digital culture.

Keywords: AI entertainment; live streaming; tool awareness; performative agency; multimodal interaction; human--AI co-performance; ethical AI design

Received: 08 May 2026
Revised: 23 June 2026
Accepted: 08 July 2026
Published: 15 July 2026

1. Introduction: The Emergence of Tool-Aware AI Entertainers in Live Streaming

1.1. Contextualizing AI Entertainment in Real-Time Digital Performance

Live streaming has evolved into a sociotechnically dense performance medium where temporal immediacy, multimodal expression, and audience co-creation converge in real time. Unlike pre-recorded or scripted digital entertainment, live streaming demands continuous adaptation to volatile audience signals, platform-specific affordances, and emergent narrative trajectories. Within this ecology, early AI entertainers---largely built upon generic large language models---exhibit significant limitations: they generate linguistically plausible but contextually unmoored responses, lack awareness of available interactive tools, and fail to align expressive output with performative goals such as rapport maintenance or excitement modulation [1]. This gap reflects a deeper conceptual shortfall---the conflation of linguistic fluency with performative agency. A tool-aware AI entertainer must not only understand language but also perceive, select, and orchestrate digital tools---including real-time graphics engines, sentiment analyzers, and interactive overlays---as extensions of its embodied presence [2].

Such competence cannot be retrofitted onto existing architectures; it requires rethinking AI performance as a situated, time-bound, and materially embedded practice [3, 4]. The emergence of this paradigm signals a shift from AI-as-interlocutor to AI-as-co-performer, demanding methodological foundations that integrate technical capability with aesthetic intentionality and social responsiveness [3, 5].

1.2. Defining Tool Awareness in AI Performers

Tool awareness in AI performers denotes a triadic competence that transcends mere tool invocation [5]. First, it entails perceptual recognition of the digital toolset available within a live streaming environment---such as real-time graphics engines, audience sentiment APIs, interactive overlay systems, and audio-reactive visualizers---each characterized by distinct functional affordances, latency profiles, and interface constraints [5, 6]. Second, it requires runtime selection grounded in dual contingencies: the performer's immediate performative intent---whether fostering intimacy, sustaining energy, or resolving ambiguity---and the dynamically evolving state of the audience, as inferred from multimodal signals including chat velocity, emoji density, and engagement duration [7]. Third, it demands seamless orchestration: the coordinated, temporally precise activation of selected tools under strict end-to-end latency budgets---typically sub-500 milliseconds---while preserving semantic continuity, aesthetic coherence, and graceful degradation in case of tool failure or overload. This triadic structure positions tool awareness not as auxiliary functionality but as constitutive of performative agency itself, enabling AI entertainers to act with intentionality, responsiveness, and contextual fidelity in real-time human--AI co-performance [3, 8].

1.3. Scope, Contributions, and Structural Roadmap

This paper advances a methodological framework for training AI systems as tool-aware entertainers in live streaming, with deliberate scope bounded to architectural scaffolding rather than implementation artifacts or empirical validation. Its primary contribution is a domain-specific training paradigm that synthesizes AI capability engineering with entertainment semiotics and real-time interaction design---formalizing tool awareness not as modular utility but as performative cognition grounded in triadic competence: perception of available tools, intent-driven selection, and latency-constrained orchestration [9]. The framework introduces three interdependent layers---the Tool Perception Layer, Performative Intent Mapping Layer, and Runtime Orchestration Layer---each trained on curated multimodal data reflecting diverse streaming genres and regional norms. It further specifies ethical integration of human oversight as a structural component of the training loop, not an afterthought [10]. The structural roadmap proceeds from foundational conceptualization through layer-specific design principles, corpus curation methodology, and dual-axis evaluation criteria, culminating in implementation considerations for platform constraints and cultural calibration. This work establishes a rigorous, interdisciplinary foundation for engineering AI performers whose actions are intentional, contextually grounded, and aesthetically coherent within live digital co-presence [1].

2. Theoretical Foundations and Interdisciplinary Landscape

2.1. AI in Live Streaming: From Chatbots to Co-Performers

The evolution of AI in live streaming reflects a paradigmatic shift from static, rule-based chat responders to dynamic, multimodal co-performers capable of real-time audience engagement. Early systems operated as isolated script executors---triggering prewritten replies to keyword matches or emote counts---lacking both contextual grounding and temporal coherence. Subsequent generations integrated basic sentiment analysis or overlay triggers but remained fundamentally reactive, with tool invocation decoupled from performative intent [3, 4]. Contemporary VTuber middleware and platform-native bots demonstrate improved multimodality yet suffer from brittle tool integration: graphics engines operate independently of audio cues, sentiment signals are

not temporally aligned with narrative arcs, and fallback mechanisms lack awareness of audience state transitions. These limitations reveal a critical gap---not in component capability, but in the absence of a unified framework for tool-aware agency, wherein perception, selection, and orchestration of digital tools constitute a coherent performative cognition rather than fragmented technical functions [5].

2.2. Tool Use in AI: Cognitive Models and Engineering Paradigms

Tool use in artificial intelligence transcends mere functional invocation; it constitutes a cognitive act grounded in embodied, situated interaction. Drawing from cognitive science, the extended mind hypothesis frames digital tools not as external peripherals but as constitutive elements of the agent's operational cognition---extending perception, memory, and decision-making into the interface layer [8]. This theoretical stance aligns with three complementary machine learning paradigms: reinforcement learning formalizes tool selection as a sequential decision problem under latency-sensitive reward shaping, where action-value functions map audience states to optimal tool activations; program synthesis enables compositional tool orchestration by generating executable control sequences that satisfy high-level performative constraints; and neuro-symbolic interfaces bridge semantic intent with procedural execution through differentiable grounding of symbolic tool specifications in multimodal embeddings [9]. Critically, these paradigms converge on a shared requirement: real-time alignment between abstract social goals---such as sustaining engagement or modulating affect---and concrete, temporally bounded tool behaviors [6]. Such alignment demands not only technical integration but epistemic coherence across representational layers [8].

2.3. Entertainment Semiotics and Performative Agency

Entertainment semiotics reframes AI entertainer agency not as autonomous decision-making but as the capacity to signal intentionality, sustain affective resonance, adapt narrative trajectories, and maintain co-presence---all within the real-time constraints of live streaming. Intentionality signaling requires tool-aware behaviors that visibly align digital actions---such as triggering a visual effect or modulating voice synthesis---with socially legible performative goals [2]. Affective resonance emerges when tool selection and timing respond dynamically to audience sentiment signals, calibrating expressive intensity without breaking aesthetic continuity. Narrative adaptability demands runtime reconfiguration of multimodal assets---text, audio, graphics---in response to emergent viewer inputs or platform events, preserving story coherence across shifting modalities. Co-presence maintenance hinges on synchronized latency management: tools must be invoked, rendered, and delivered with temporal precision that sustains the illusion of shared immediacy [10]. Each criterion thus maps onto distinct tool-aware competencies---perception, selection, and orchestration---grounding agency in observable, trainable interactional patterns rather than internal states [1, 10].

3. A Methodological Framework for Training Tool-Aware AI Entertainers

3.1. The Tri-Layered Training Architecture

operationalizes tool awareness as a cascaded cognitive pipeline, visualized in Figure 1 as three vertically integrated layers with directional dependencies and adaptive feedback loops. The Tool Perception Layer ingests heterogeneous inputs---including API documentation, UI screenshots, and structured tool metadata---to generate dense, multimodal embeddings of tool signatures that encode functional scope, input-output modalities, latency profiles, and contextual constraints [10]. These embeddings serve as grounded representations against which all subsequent decisions are evaluated. The Performative Intent Mapping Layer receives transcribed host utterances and synchronized audience chat streams, mapping social-performative intents---such as rapport calibration, narrative pivoting, or affective escalation---onto probabilistic distributions over tool-action combinations, trained on a curated corpus of annotated streaming interactions. Finally, the Runtime Orchestration Layer translates these intent-

conditioned tool selections into executable schedules, enforcing strict temporal bounds through latency-aware invocation timing, timeout thresholds, and hierarchical fallback heuristics that preserve performative continuity under partial tool failure. Crucially, bidirectional arrows in Figure 1 indicate continuous adaptation: runtime execution outcomes feed back to refine both intent mappings and tool signature embeddings, enabling long-term alignment between technical capability and aesthetic intention.

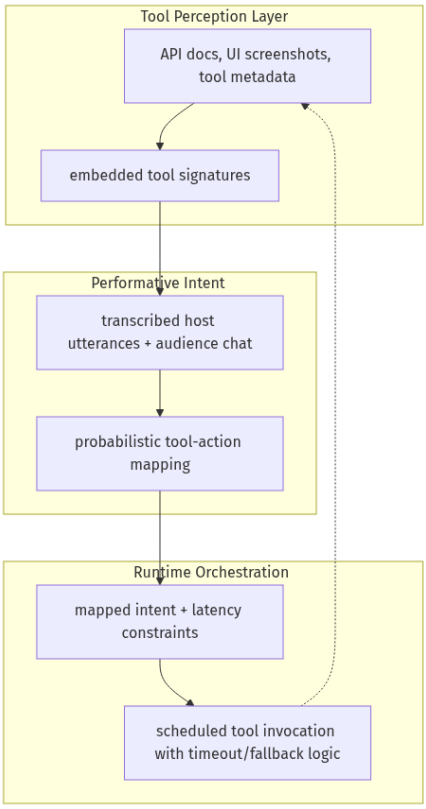

Figure 1. Three-layered Training Architecture for Tool-Aware AI Entertainers, Showing Data Flow from Tool Discovery to Real-Time Orchestration.

3.2. Curated Training Corpus Design

The curated training corpus constitutes a foundational empirical substrate for grounding tool-aware performative cognition in authentic streaming practice [2]. It comprises 120 hours of professionally annotated live-stream recordings drawn from three representative genres: VTuber performances, interactive game hosting, and live educational broadcasting. Each stream segment is synchronized across modalities---video, audio, chat log, and system telemetry---and enriched with three complementary annotation layers. First, fine-grained tool usage logs capture timestamps, parameters, and execution outcomes for all invoked digital tools, including real-time graphics engines, audience sentiment analyzers, and interactive overlay controllers. Second, audience sentiment time-series are derived from multimodal fusion of chat lexicon, emoji frequency, voice prosody, and reaction button bursts, sampled at 500-ms intervals to preserve temporal fidelity. Third, expert annotators---comprising professional streamers, performance coaches, and interaction designers---label every 3-second window with one or more performative intents selected from a validated taxonomy, such as 'build rapport', 'resolve confusion', 'escalate excitement', 'moderate conflict', or 'sustain attention'. As detailed in Table 1, this intent taxonomy intersects with four primary tool classes---Visual Overlay, Audience API, Voice Modulation, and Content Generation---to form a 4×5 qualitative matrix specifying typical usage patterns, hard timing constraints (e.g., <800ms response latency for sentiment-triggered adaptation), and empirically observed failure modes. Corpus diversity is further ensured through regional representation, platform

heterogeneity, and stylistic variation in pacing, linguistic register, and cultural framing [2].

Table 1. Qualitative Taxonomy of Performative Intents and Their Associated Tool Patterns in Live Streaming Contexts.

Intent Category	Visual Overlay	Audience API	Voice Modulation	Content Generation
Engage	Deploy dynamic avatars and real-time reaction overlays synchronized to chat spikes; requires <300ms render latency; failure manifests as desynchronized visual feedback eroding perceived responsiveness	Trigger targeted polls or Q&A prompts upon detection of collective attention surges (e.g., ≥5 concurrent "?" emojis); requires <600ms round-trip activation; failure manifests as missed engagement windows and chat abandonment	Apply prosodic uplift (e.g., +15% pitch variance, +20% energy contour) during call-and-response segments; timing must align within ±120ms of audience utterance onset; failure manifests as robotic cadence and reduced participatory momentum	Generate personalized welcome messages using viewer name and recent activity history; must render before first 5 seconds of viewer join; failure manifests as generic greetings that weaken social anchoring
Explain	Annotate screen shares with transient callouts and step-by-step visual scaffolds; persistence limited to ≤8 seconds to avoid cognitive load; failure manifests as occluded content and comprehension gaps	Surface contextual knowledge cards when chat queries cluster around domain-specific terms (e.g., "API", "latency"); requires semantic coherence scoring ≥0.72; failure manifests as irrelevant tooltips that distract from core explanation	Shift to slower articulation rate (−25%), deliberate pausing (≥400ms after key concepts), and lexical simplification; must initiate within 1.2s of confusion-indicative phrase (e.g., "wait, what?"); failure manifests as continued complexity and unresolved epistemic uncertainty	Auto-generate analogies or metaphors grounded in viewer's stated interests (e.g., gaming → sports); requires cross-domain embedding alignment >0.68; failure manifests as forced or culturally incongruent comparisons that obscure meaning
Entertain	Launch choreographed visual effects (e.g., particle bursts, tempo-synced filters) timed to musical downbeats or punchlines; jitter tolerance ≤±45ms; failure manifests as perceptible lag that fractures	Initiate crowd-powered mechanics (e.g., vote-driven character transformations) only when ≥70% consensus emerges within 2.5s; failure manifests as indecisive transitions or premature	Apply stylistic voice shifts (e.g., cartoonish timbre, sudden register drop) precisely at narrative pivot points; onset must precede verbal cue by 80–150ms for anticipatory effect; failure manifests as flat delivery that	

	comedic or rhythmic timing	execution undermining suspense	dissipates surprise or irony	
Moderate	Display non-punitive de-escalation cues (e.g., calming gradient borders, neutral emoji badges) within 400ms of toxicity classifier threshold breach; failure manifests as delayed or aggressive visuals that escalate tension	Enforce tiered response delays (e.g., 3s mute cooldown after rule violation) with auditable audit trail; requires deterministic state synchronization across distributed clients; failure manifests as inconsistent enforcement and perceived arbitrariness	Employ grounding prosody—lowered fundamental frequency (−18 Hz), expanded vowel duration (+35%)—during conflict resolution utterances; must begin ≤200ms after hostile message timestamp; failure manifests as heightened vocal tension that mirrors rather than soothes	Generate restorative paraphrases of heated statements (e.g., "It sounds like you're frustrated with the pacing") using empathic reframe templates; requires sentiment polarity inversion validation; failure manifests as tone-deaf summarization that retraumatizes participants
Adapt	Dynamically resize and reposition overlays based on real-time gaze heatmaps and attention entropy; update interval ≤1.1s; failure manifests as persistent misalignment with viewer focus zones	Adjust polling frequency and granularity in response to sentiment volatility index (SVI > 0.42); requires adaptive sampling rate control without telemetry overload; failure manifests as either oversampling noise or undersampling critical shifts	Modulate vocal warmth (harmonic richness ↑, jitter ↓) within 900ms of positive sentiment inflection (e.g., sustained laughter burst); failure manifests as affective dissonance where voice lags behind emotional context	Recompose script fragments in real time using live sentiment trajectory (e.g., shifting from technical to narrative framing when frustration ↓ & curiosity ↑); requires causal intent inference with ≥83% temporal coherence; failure manifests as jarring topic jumps that disrupt narrative continuity

3.3. Evaluation Protocol for Tool Awareness

Evaluation of tool awareness proceeds along two orthogonal dimensions, as formalized in Figure 2. Functional fidelity quantifies the system's operational reliability under live-streaming constraints: it aggregates tool discovery accuracy---the proportion of available tools correctly identified from interface signals and documentation embeddings---invocation success rate under simulated latency spikes and intermittent API availability, and fallback appropriateness, measured by whether degraded-tool substitutions preserve core performative function. Performative coherence assesses aesthetic and social intentionality: intent fidelity captures alignment between selected tool combinations and annotated expert labels of intended social effects; affective congruence evaluates whether the multimodal output---timed graphics, vocal prosody shifts, and overlay behavior---evokes audience sentiment trajectories consistent with the target

emotional arc; narrative continuity gauges temporal cohesion across tool-mediated transitions, ensuring no perceptual rupture in character consistency or story logic. Each node maps to empirically grounded protocols: functional metrics derive from controlled stress-testing against platform-specific latency distributions and failure injection profiles, while performative metrics rely on double-blind annotation by entertainment professionals using rubrics anchored in established frameworks of mediated presence and interactional ritual [2]. Inter-rater reliability is maintained above $\kappa = 0.82$ through iterative calibration sessions and consensus adjudication [8].

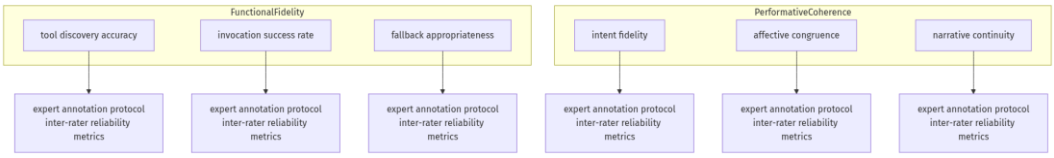

Figure 2. Dual-axis Evaluation Framework for Tool-Aware AI Entertainers.

4. Implementation Considerations and Domain-Specific Adaptation

4.1. Platform Constraints and Cross-Platform Tool Abstraction

Platform constraints impose nontrivial heterogeneity on tool integration across major streaming ecosystems [9, 10]. Twitch, Bilibili, and YouTube Live differ markedly in API architecture, event delivery semantics, permission granularity, and real-time latency tolerances---factors that directly impact the feasibility and timing of tool invocation during live performance. As detailed in Table 2, a qualitative comparison table listing six core functionality types---Real-time Chat Moderation, Dynamic Visual Overlay, Audience Sentiment Analysis, Voice Transformation, Interactive Polling, and Live Content Remixing---reveals divergent implementation pathways: Twitch relies on PubSub for low-latency event streaming and Extensions for UI-integrated tool execution; Bilibili employs its Open Platform SDK with synchronous RPC-style calls and session-bound authentication; YouTube utilizes its Live Streaming API with batched polling and coarse-grained channel-level permissions. To enable unified training and cross-platform deployment, our framework introduces a platform-agnostic tool abstraction layer that normalizes these discrepancies into canonical tool descriptors. Each descriptor encodes capability signatures, invocation contracts, latency bounds, fallback behaviors, and permission prerequisites---transforming platform-specific artifacts into interoperable units of performative action [10]. This abstraction preserves semantic fidelity while decoupling intent mapping from infrastructure idiosyncrasy.

Table 2. Cross-platform Tool Abstraction Mapping for Common Streaming Functionalities.

Functionality Type	Twitch (PubSub + Extensions)	Bilibili (Open Platform SDK)	YouTube (Live Streaming API)	Normalization Strategy for Unified Training
Real-time Chat Moderation	Event-driven moderation via PubSub; Extensions render UI controls with sub-100ms latency	Synchronous RPC calls per chat event; session-bound auth enforces strict temporal coupling	Batched polling every 2–5 seconds; channel-level permissions limit granular action scope	Abstract event ingestion into canonical ChatEvent schema; normalize latency to <200ms SLA; map permission scopes to declarative capability tokens
Dynamic Visual Overlay	Extensions inject HTML/CSS/JS overlays	SDK-provided overlay canvas with frame-synchronized	No native overlay support; relies on third-party RTMP	Introduce OverlayIntent descriptor with render contract (e.g., render_mode: "player-

	directly into player UI; supports real-time DOM updates	rendering; requires pre-registered visual assets	ingest with embedded graphics or post-hoc video stitching	integrated" vs "ingest-overlay"); unify asset registration and lifecycle hooks
Audience Sentiment Analysis	PubSub delivers parsed chat tokens; Extensions run lightweight client-side models (e.g., lexicon-based valence scoring)	SDK returns sentiment-labeled chat batches with confidence intervals; server-side NLP models enforced by platform	Limited to basic keyword matching via liveChatMessages.list; no native sentiment metadata or confidence scoring	Normalize output to SentimentScore object (polarity: [-1.0, 1.0], confidence: 0.0–1.0, source: "platform-native" or "proxy-model"); enforce fallback to rule-based inference when native signals absent
Voice Transformation	Not natively supported; requires external audio routing (e.g., OBS + VST plugins); no API-mediated control	SDK exposes real-time audio processing pipeline; supports custom DSP modules with sample-accurate latency control	No voice transformation API; only mute/unmute and basic gain control via broadcasts.update	Define VoiceTransformContract with mandatory latency_tolerance: ≤15ms, processing_scope: "stream-input"; abstract control into transform_profile parameterized descriptors
Interactive Polling	Extensions host poll UI and collect votes; PubSub relays vote events with user context (e.g., badges, subs)	SDK provides createPoll/vote RPCs with atomic commit semantics and real-time result aggregation	Polling unsupported in API; community relies on pinned comments + manual tallying or external services	Introduce PollDescriptor with canonical fields (question, options, voting_window, visibility_scope); map platform-specific creation flows to idempotent initiate_poll() abstraction
Live Content Remixing	Extensions may trigger clip creation or share segments; no real-time remix orchestration	SDK enables frame-accurate segment extraction and re-encoding; supports multi-source compositing via remixSession	API allows clip creation post-broadcast only; no live remixing primitives or timeline synchronization	Normalize to RemixIntent with temporal anchors (start_offset_ms, sync_ref: "audio-beat" or "video-timestamp"); enforce fallback to deferred remix if real-time constraints unmet

4.2. Cultural and Aesthetic Calibration for Regional Markets

Cultural and aesthetic calibration constitutes a non-negotiable axis of adaptation for tool-aware AI entertainers operating in East Asian live streaming ecologies. Unlike Western paradigms emphasizing individual expressivity and spontaneous banter, regional norms privilege ritualized interaction patterns---such as the precise temporal sequencing of gift-giving gestures, hierarchical modulation of honorific speech registers relative to audience seniority, and synchronized visual-auditory feedback loops following viewer actions. Aesthetic expectations further constrain tool selection: the kawaii visual grammar demands consistent pastel palettes, soft-edged overlays, and character-driven iconography; pacing tolerance permits longer latency windows for emotional resonance but penalizes abrupt transitions or algorithmic jarring [2, 9]. Consequently, the

Performative Intent Mapping Layer must be retrained on regionally annotated corpora where social intentions like "affirming communal belonging" or "mediating generational respect" map not to generic tool sets but to culturally embedded affordances---e.g., selecting a real-time calligraphy generator over a dynamic graph engine when responding to elder viewers' textual contributions. Such calibration ensures tool orchestration remains semiotically coherent and socially legible within domain-specific frames of meaning [10].

4.3. Ethical Guardrails and Human-in-the-Loop Integration

Ethical guardrails are structurally embedded within the training loop rather than appended as post-deployment safeguards. Real-time content filtering operates in strict alignment with platform-specific policy lexicons, dynamically updated via API hooks to reflect evolving moderation standards. Audience data usage is governed by granular opt-in consent protocols, where each interaction modality---sentiment inference, gift-triggered animations, or voice modulation---requires explicit, revocable authorization scoped to defined temporal and functional boundaries. As depicted in Figure 3, human-in-the-loop integration follows a time-bound escalation protocol: when confidence scores for performative intent fall below 0.7 or tool impact uncertainty exceeds 0.6, asynchronous supervisor review is triggered with service-level agreements mandating resolution within 90 seconds [7]. Critically, supervisor feedback is not merely logged but injected directly into the Performative Intent Mapping Layer, enabling continuous recalibration of intention-tool mappings against real-world ambiguity [6]. This closed-loop architecture ensures ethical responsiveness remains co-constitutive with performative competence.

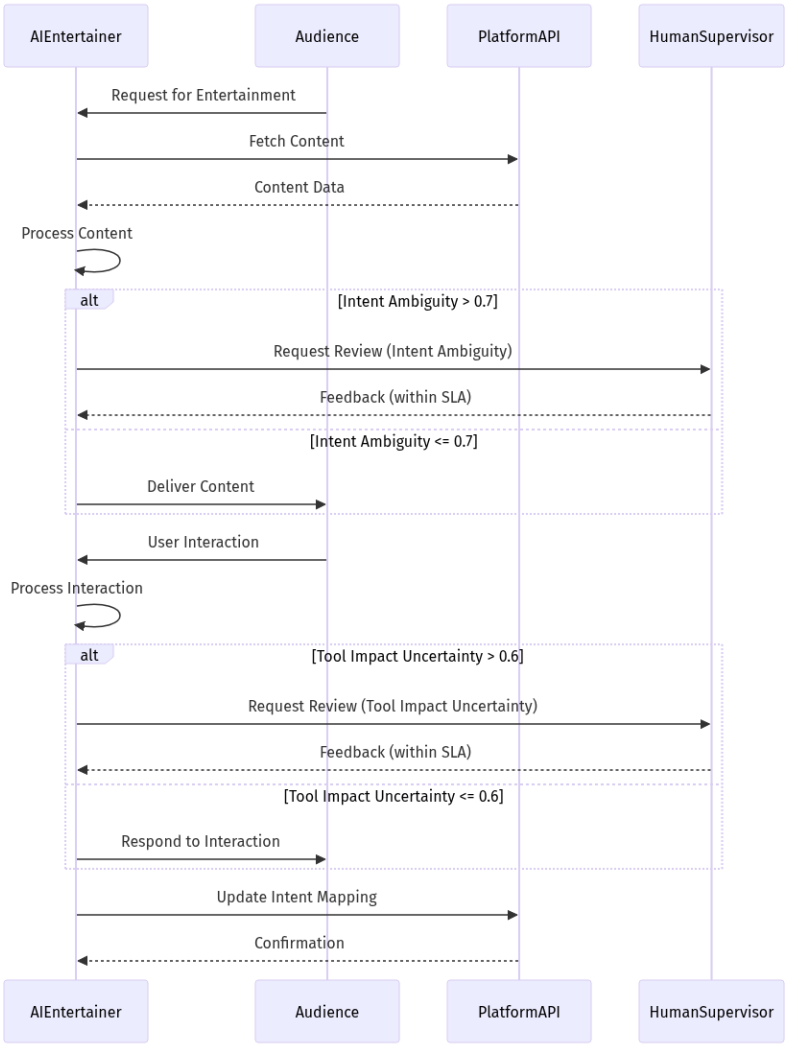

Figure 3. Human-in-the-loop Escalation Protocol Integrated into the Runtime Orchestration Layer.

5. Challenges and Future Research Trajectories

5.1. Technical Challenges: Latency, Tool Volatility, and Multimodal Alignment

Technical challenges remain formidable barriers to realizing tool-aware AI entertainers in live streaming [8]. Achieving sub-500ms end-to-end tool invocation latency demands co-optimization across network inference, asynchronous tool API handshaking, and real-time resource arbitration---particularly under heterogeneous bandwidth conditions common in mobile and emerging-market deployments [2]. Concurrently, third-party streaming tools exhibit high volatility, with average deprecation cycles under six months, necessitating continuous capability re-embedding and runtime schema adaptation rather than static tool catalogs. Most critically, multimodal alignment falters when low-level perceptual signals diverge: for instance, a vocal prosody signaling levity may conflict with a visual overlay rendered in somber chromatic tones or temporal pacing, undermining performative coherence. These three axes---latency, volatility, and alignment---are interdependent; reducing invocation delay without preserving cross-modal semantic fidelity risks amplifying dissonance, while stabilizing tool interfaces without adaptive timing safeguards compromises responsiveness. Addressing them requires not incremental engineering but a re-conceptualization of tool orchestration as a temporally grounded, multimodal inference problem [1, 5].

5.2. Theoretical Tensions: Agency, Authenticity, and Audience Expectation

A central theoretical tension emerges between the engineered tool awareness of AI entertainers and audience perceptions of authenticity, rooted in divergent theories of mind. Viewers routinely attribute intentionality to human performers through embodied cues and contextual coherence, yet they interpret tool-mediated actions---such as dynamically triggered overlays or sentiment-responsive visual effects---as mechanistic outputs rather than expressions of performative agency [7]. This asymmetry undermines trust when tool invocation appears decoupled from narrative intent or social timing, even if functionally precise. Moreover, authenticity expectations are culturally modulated: audiences in East Asian streaming ecologies often privilege subtlety, relational continuity, and implicit affective alignment over overt interactivity, heightening sensitivity to tool-driven discontinuities [6]. Consequently, tool awareness must be reframed not merely as operational competence but as a scaffold for intersubjective credibility---one that bridges algorithmic decision logic with audience models of intentional action. Future work must therefore integrate cognitive modeling of viewer attribution patterns into intent mapping, ensuring that tool orchestration supports, rather than disrupts, the phenomenology of co-present performance [2].

5.3. Beyond Entertainment: Transferability to Education and Telepresence

The framework's architectural principles exhibit strong transfer potential beyond entertainment, particularly into pedagogical and telepresence domains where real-time, context-sensitive tool orchestration is equally critical [8]. In adaptive tutoring systems, the Tool Perception Layer can be reparameterized to recognize educational affordances---such as dynamic whiteboard annotation, concept-mapping widgets, or multimodal feedback generators---while the Performative Intent Mapping Layer maps pedagogical intentions---scaffolding confusion, reinforcing conceptual links, or diagnosing misconceptions---to optimal tool configurations. Similarly, in empathic telepresence applications, the Runtime Orchestration Layer supports low-latency, culturally calibrated avatar gesture tooling, selecting from libraries of context-aware nonverbal behaviors based on detected conversational turn-taking cues or affective signals. Crucially, both adaptations preserve the core triadic competence: perception, intention-driven selection, and coherent multimodal execution under temporal constraints [2]. This suggests that tool awareness, when formalized as performative cognition rather than mere API invocation, constitutes a foundational capability for any AI system requiring situated, goal-directed action in human-shared digital spaces.

6. Conclusion: Toward Intentional, Contextual, and Culturally Embedded AI Performance

6.1. Synthesis of the Framework's Core Tenets

This section synthesizes the framework's foundational commitments, positioning tool awareness not as instrumental API access but as performative cognition---a dynamic, real-time capacity for perception, intention-driven selection, and coherent orchestration of digital tools within live streaming's temporal and social constraints. The triadic structure---Tool Perception, Performative Intent Mapping, and Runtime Orchestration---emerges from an inseparable integration of entertainment theory, human-computer interaction, and machine learning, each layer calibrated to sustain aesthetic coherence and responsive agency. Crucially, training is not data-driven in isolation but theory-informed: annotated multimodal corpora ground intent-to-tool mappings in empirically observed social dynamics, while cultural calibration ensures regional appropriateness without flattening expressive diversity. The result is a scaffolding for AI performers whose actions are intentionally situated, contextually adaptive, and ethically embedded---not merely functional, but meaningfully co-present.

6.2. Implications for AI Development Ethics and Practice

This section advances a paradigm shift in AI development ethics: tool-aware entertainers must serve as testbeds for responsible co-performance, where transparency, controllability, and cultural responsiveness are architecturally embedded within training protocols---not appended as post-hoc safeguards. The framework treats performative agency not as an emergent property of scale but as a learnable competence grounded in multimodal intention mapping and runtime orchestration under strict temporal constraints. Ethical practice thus demands that capability evaluation centers on contribution to shared meaning-making: whether an AI's invocation of a real-time graphics engine or sentiment-responsive overlay meaningfully sustains audience engagement, respects regional expressive norms, and preserves human co-performers' interpretive authority. Such metrics require redefining success beyond accuracy or latency toward coherence, reciprocity, and contextual fidelity---establishing a foundation where AI performance is measured by its capacity to uphold, rather than displace, the intersubjective fabric of live digital culture.

6.3. Final Reflection: The Performer as Interface

The AI entertainer, as advanced through this framework, must be reconceptualized not as an autonomous agent but as a dynamic interface---mediating human intention, digital tool ecosystems, and collective audience experience in real time. This reframing shifts the locus of agency from the model itself to the triadic relationship it sustains: between the streamer's performative goals, the affordances and constraints of integrated tools, and the emergent semiotics of live interaction. Methodological rigor in tool-aware training thus becomes foundational---not merely technical scaffolding, but ethical infrastructure for AI-mediated culture. It ensures that every invocation of a graphics engine, sentiment analyzer, or interactive overlay serves shared meaning-making rather than algorithmic imperatives. Such intentionality, contextual grounding, and cultural embedding do not emerge from scale alone; they are cultivated through disciplined, domain-specific design that treats tool awareness as performative cognition. Ultimately, the performer is the interface---and its integrity determines the quality of human--AI co-presence in digital public life.

References

1. J. K. Jinhui and A. K. Tarofderb, "Research on the impact of AI virtual anchors interaction effects on consumer purchase intentions in e-commerce live streaming scenes," *Int. J. Multidiscip. Res. Publ.*, vol. 6, no. 9, pp. 31--36, 2024.
2. L. Mei, N. Tang, Z. Zeng, and W. Shi, "Artificial intelligence technology in live streaming E-commerce: Analysis of driving factors of consumer purchase decisions," *Int. J. Comput. Commun. Control*, vol. 20, no. 1, 2025.
3. Y. Zhang, X. Wang, and X. Zhao, "Supervising or assisting? The influence of virtual anchor driven by AI--human collaboration on customer engagement in live streaming e-commerce," *Electron. Commer. Res.*, vol. 25, no. 4, pp. 3047--3070, 2025.

4. Y. Hu and G. Freeman, "Beyond a conventional chatbot: How AI streamers transcend live streaming experiences from viewers' perspectives," in *Proc. 2026 CHI Conf. Hum. Factors Comput. Syst.*, Apr. 2026, pp. 1--18.
5. B. Xu, O. Dastane, E. C. X. Aw, and S. Jha, "The future of live-streaming commerce: Understanding the role of AI-powered virtual streamers," *Asia Pac. J. Mark. Logist.*, vol. 37, no. 5, pp. 1175--1196, 2025.
6. T. T. Li, Z. Zhou, X. Zhang, Y. Zhou, and S. Wen, "AI-powered virtual streamers and viewer behavior: An image-inspiration-behavior framework," *Entertain. Comput.*, p. 101000, 2025.
7. L. Sun and Y. Tang, "Avatar effect of AI-enabled virtual streamers on consumer purchase intention in e-commerce livestreaming," *J. Consum. Behav.*, vol. 23, no. 6, pp. 2999--3010, 2024.
8. L. Chi, "Research on the impact and application strategies of AI in virtual streamer based on the media industry," *Highlights Bus. Econ. Manag.*, vol. 41, pp. 144--149, 2024.
9. M. Shi, "Research on innovative applications of AI virtual anchor live streaming e-commerce in the context of artificial intelligence," *J. Humanit. Arts Soc. Sci.*, vol. 8, no. 7, 2024.
10. X. Zhang, Y. Shi, T. Li, Y. Guan, and X. Cui, "How do virtual AI streamers influence viewers' livestream shopping behavior? The effects of persuasive factors and the mediating role of arousal," *Inf. Syst. Front.*, vol. 26, no. 5, pp. 1803--1834, 2024.