Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment
(AIIDE 2026)

Finding Moments Worth Remembering: Multimodal Evidence and Multi-Stage Reasoning for Illustrating Actual-Play TTRPG Sessions

Gianlucca Lodron Zuin[1], Túlio Corrêa Loures[1], Eduardo de Moura[1], Adriano Alonso Veloso[1,2], Marcelo Coelho[3]

[1]Kunumi Institute
[2]Universidade Federal de Minas Gerais
[3]Massachusetts Institute of Technology
{gianlucca, tulio, eduardo.moura, adriano}@kunumi.com, marceloc@mit.edu

Abstract

Tabletop role-playing games present a challenging domain for artificial intelligence, particularly for reasoning about emotional shifts and narrative significance in long-form actual-play sessions. Players improvise shared stories over hours of dialogue, performance, rules negotiation, and social reaction, while the significance of memorable moments may depend on context beyond the local passage. We propose a hierarchical multimodal architecture for selecting emotionally and narratively significant moments as candidates for visual representation. The approach extracts speech, speaker, affective, conversational, and table-response evidence, contextualizes these signals through short-, medium-, and long-term narrative memory, and uses a multi-stage LLM-based judge to evaluate local, scene, and memory relevance before visualization. We evaluate moment selection using silver labels derived from viewer signals, and assess generated visual memories through human judgments of image–audio alignment. The system recovers referenced timestamps above a chance baseline at exact-window resolution, indicating mild alignment with viewer-perceived salience. Human evaluation indicates that the pipeline produces grounded images for well-chosen moments and that these images are preferred over those generated directly from transcripts. Together, these results represent a first step toward memory-supported visual externalization in long-form collaborative storytelling.

Code — https://doi.org/10.6084/m9.figshare.33372268

Introduction

Tabletop Role-Playing Games (TTRPGs, or RPGs, for short) combine group interaction, subjective rule application and adjudication, improvised collaborative storytelling, and character acting. Typically, these games are played with one person filling the role of the Game Master (GM), while the other participants are players who mostly roleplay as a single character. As the player characters interact with each other and with the game elements, by means of narrative actions and supported by rules, the players are able to collectively build a story and continuously alter the state of the game world. These game interactions can take place and have implications for the duration of short scenes, hours-long play sessions, or even over campaigns that span years.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Ellis and Hendler (2017) discuss TTRPGs as human-oriented games that represent great challenges for AI researchers. One such problem is that of measuring and enhancing player immersion. For the collective interactions to work, players need to stay involved with the game: getting into character and keeping track of the ever-evolving state of a fictional world places sustained demands on their attention, (Kellogg et al. 2013; Jennett et al. 2008) and, as Brown and Cairns (2004) discuss, to achieve the deepest level of immersion, one of the main barriers is atmosphere.

Art, music, props, and other sensory stimuli are often used to set the atmosphere of a game and increase the engagement of the players (Borecky 2021; Wizards of the Coast 2014; Petersen et al. 2016). Even objects made for practical use during the game, such as the *GM screen*, often feature artwork on the side facing the players to evoke some of the game's themes and setting. In practice, however, these traditional enhancements are typically selected as non-campaign-specific elements or prepared in advance by the GM (Acharya, Mateas, and Wardrip-Fruin 2021). These game session enhancements are not guaranteed to be tailored to the specific narrative events that dynamically unfold during play, and may not always adequately support the game scenes as the emotional register of play shifts.

In games played digitally, or with the assistance of digital devices, such elements can instead be searched for, or generated, during the session itself. However, finding the right piece of art, or composing an appropriate prompt for what is desired, takes time and attention that is drawn away from the game itself (Ferreira, Lelis, and Whitehead 2020). A system that dynamically identifies the right moments to intervene, generating mood-setting elements without explicit prompting, could enhance the atmosphere of such games while minimizing the effort demanded of the GM.

With that intention, we want to know if we can correctly identify the ideal moments for automatic intervention. And, given those selected moments, if we can generate scene descriptions that can be used to prompt images that fit well with the respective game events. We propose a hierarchical multimodal architecture that operates directly on digital audio recordings of actual-play sessions, as exemplified by Figure 1. From the recording, the system derives time-aligned speech, speaker, affective, and conversational evidence; it contextualizes this evidence through

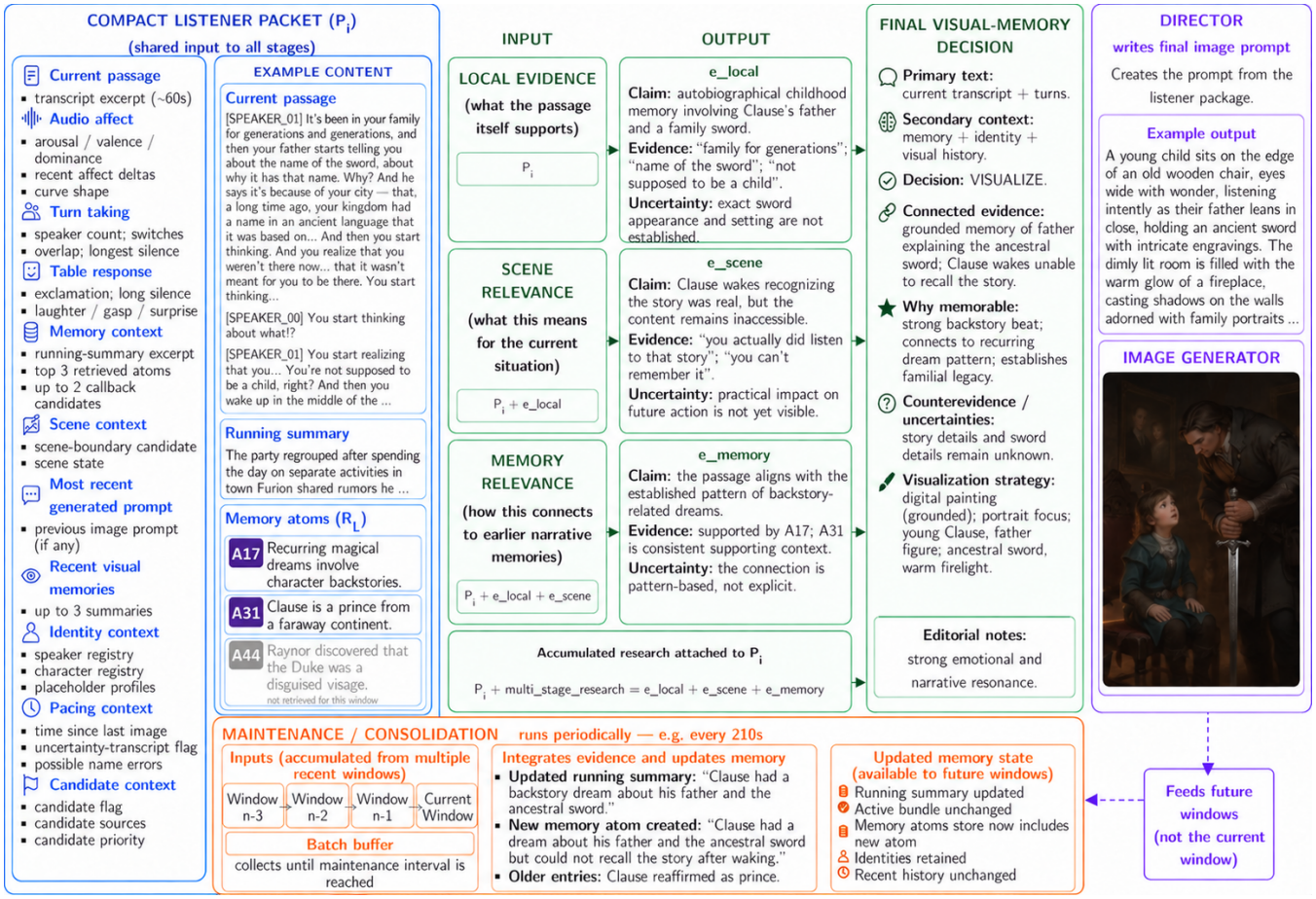


Figure 1: Example of narrative evaluation and visualization. A player character's dream is evaluated using local, scene, and memory context, selected for its exposition and narrative relevance, and subsequently visualized.

short-, medium-, and long-term narrative memory; it applies a multi-stage LLM-based listener that decides, for each candidate window, whether the moment should be externalized as a *visual memory* and, if so, how it should be staged for image generation. Our contributions are as follows:

- We formulate visual externalization in long-form actual-play TTRPGs as a moment-selection problem over decision windows, coupling narrative-memory context with a multi-stage selection procedure.
- We describe a hierarchical multimodal memory that aligns heterogeneous audio evidence to decision windows and maintains retrievable narrative context across a multi-hour session.
- We evaluate moment selection on labels derived from viewer engagement on a public actual-play, reporting performance as lift over a coverage-matched baseline.

Related Work

Tabletop role-playing games pose a demanding challenge for AI because they combine open-ended natural language action, improvised multi-party narration, hidden or partially shared state, and subjective rule adjudication (Ellis and Hendler 2017). The difficulty is not only that the interaction is open-ended, but that role-playing itself is a social and narrative practice. Role-playing game studies have long treated TTRPGs as more than formal rule systems, emphasizing social dynamics and collaborative production (Fine 2002; Cover 2014). More broadly, RPGs have been positioned as interdisciplinary objects at the intersection of game studies, narrative, performance, media, and culture (Deterding and Zagal 2018). In actual-play recordings, this means that the significance of a moment may depend on shared history, social context, and the evolving fictional situation, rather than only local utterance.

This social and narrative complexity has also motivated computational work on RPGs as dialogue, summarization, and state tracking (Rameshkumar and Bailey 2020; Callison-Burch et al. 2022; Zhu et al. 2023a). Together, these works show the importance of reasoning over long dialogue histories and evolving fictional state. Other systems assist specific aspects of play, such as LLM-generated prose, summaries, and brainstorming assistance (Zhu et al. 2023b), encounter design optimization for difficulty and player contribution (Shyne and Cooper 2024), and game-master-like decision-making in complex tabletop board-game environments (Best, Lucas, and Gaina 2024).

In video analysis, selecting engaging moments from long recordings is commonly studied as highlight detection. Early work learned domain-specific highlight rankers by using edited videos as weak supervision (Sun, Farhadi, and Seitz 2014). More recent benchmarks scale this idea with explicit salience annotations or viewer-behavior signals (Lei, Berg, and Bansal 2021; Sul, Han, and Lee 2023). Audience reactions have also been used directly: Fu et al. (2017) predict highlights in e-sports streams using synchronous chat reactions together with visual features. The silver labels in this study follow this audience-driven tradition, treating timestamped viewer comments and engagement heatmaps as noisy proxies for the true highlight labels.

The evidence layer in this work builds on affective computing. We use dimensional descriptions of affect, such as valence and arousal, following the circumplex model (Russell 1980). Modern speech emotion recognition provides tools for estimating affective cues from audio, including transformer-based and self-supervised models (Wagner et al. 2023; Ma et al. 2024). Recent work in speech emotion recognition has also begun to combine acoustic, semantic, and contextual evidence before making an affective decision. Zhao et al. (2025) propose C^{2}SER, which integrates Whisper-based semantic perception, Emotion2Vec-S acoustic perception, and staged contextual reasoning to improve the stability of speech emotion recognition. This provides a useful precedent for our evidence design, though our work targets a different downstream task: C^{2}SER infers emotion labels, whereas our listener uses multiple types of evidence to decide whether a passage is narratively meaningful and visually groundable. Affective video analysis has also long studied the automatic representation and modeling of emotional content in video (Hanjalic and Xu 2005), with later surveys distinguishing direct audiovisual approaches from those based on viewer responses (Wang and Ji 2015).

The systems closest to our deployment context in terms of emotion and engagement are those that adjust ambience based on player speech. Bardo selects background music in real time based on emotions extracted from transcribed player speech (Padovani, Ferreira, and Lelis 2017; Ferreira, Lelis, and Whitehead 2020). The immersive role of sound in D&D has also been discussed in ludomusicology (Borecky 2021). Prior work on narrative emotion shows that stories can be modeled as emotional trajectories rather than isolated affective labels. Reagan et al. (2016) identify recurring emotional arcs across a large corpus of fiction, and sentence-level models have been proposed for predicting emotion across narrative stories (Too et al. 2023). These approaches are typically applied to completed texts or fixed narrative sequences. The broader notion of narrative salience is also related to work on narrative engagement (Busselle and Bilandzic 2009). This perspective suggests that memorable actual-play moments need not be the loudest or most emotionally intense moments. They may also be moments that focus attention, clarify stakes, or deepen the sense of presence in the fictional situation.

Regarding story visualization and generated visual memories, works such as StoryGAN and StoryDALL-E (Li et al. 2019; Maharana, Hannan, and Bansal 2022) explore the complexities of generating images from multi-scene and multi-character narratives. The term *visual memory* also has a cognitive motivation, based on dual-coding theory (Paivio 1990). In our setting, the generated image is therefore not a replacement for the verbal event, but its visual counterpart.

Our work does not propose a new text-to-image model. Instead, it studies the upstream problem of deciding which moments from long-form play should be visualized and how those moments should be grounded in the surrounding narrative context. The generated image is therefore used as an evaluation endpoint, rather than as the sole contribution.

Method

We model a long-form actual-play tabletop role-playing game session as an audio recording from which time-aligned narrative evidence is derived. The method evaluates short decision windows and selects a subset for visual externalization. Selected windows must satisfy three requirements. First, they must be supported by local acoustic, linguistic, and conversational evidence. Second, they must matter within the evolving fictional state. Finally, they must be grounded enough to support a visual representation. We refer to the resulting artifacts as *visual memories*. A visual memory is not merely an image, but a structured record containing the source interval, supporting evidence, retrieved narrative context, editorial decision, grounding constraints, prompt package, and rendered image.

The architecture, illustrated in Figure 2, follows this definition by separating evidence construction, listener-based moment selection, and grounded visual externalization. We first define the session representation and decision units. We then describe how compact listener packets are built from local evidence, scene state, retrieved memory, identity context, and recent visual history. Finally, we formalize the multi-

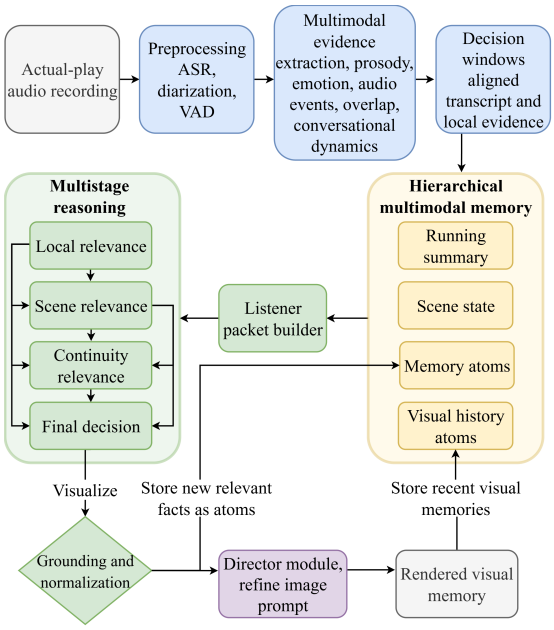


Figure 2: Pipeline of the proposed approach.

stage listener as a sequence of calls over the same packet, where later calls receive the evidence records produced by earlier calls. Let the primary input be an audio recording A of duration L. From A, the preprocessing stage derives an aligned interaction trace:

$$X = \{x_j\}_{j=1}^{J}.$$

Here J is the number of extracted interaction units. Each unit is represented as:

$$x_j = (u_j, s_j, \tau_j^0, \tau_j^1, \eta_j),$$

where u_j is a transcribed utterance or conversational event, s_j is a diarized speaker label, $(\tau_j^0, \tau_j^1) \subset [0, L]$ are temporal boundaries, and η_j contains derived interaction cues such as silence, overlap, voice activity, or short reactions.

The session is evaluated through a finite set of decision windows, where each window is a time interval:

$$\mathcal{W} = \{W_i\}_{i=1}^{K}, \qquad W_i = [t_i^0, t_i^1] \subset [0, L].$$

We use one-minute windows based on preliminary experiments on a separate actual-play campaign, balancing sufficient narrative and acoustic context against the risk of merging multiple sequential actions. The value $K = |\mathcal{W}|$ is the number of decision windows evaluated for a session. For each window, the aligned transcript segment is:

$$T_i = \{x_j \in X \mid [\tau_j^0, \tau_j^1] \cap W_i \neq \emptyset\}.$$

At decision step i, the system maintains a memory and continuity state:

$$\mathcal{M}_{i-1} = \left(M_{i-1}^{run}, M_{i-1}^{scene}, M_{i-1}^{atom}, M_{i-1}^{active}, M_{i-1}^{id}, M_{i-1}^{vis}\right).$$

Here M_{i-1}^{run} is the running session summary, M_{i-1}^{scene} is the current scene state, M_{i-1}^{atom} is the set of retrievable memory atoms, M_{i-1}^{active} is the active memory bundle, M_{i-1}^{id} stores speaker and character registries, and M_{i-1}^{vis} stores previously committed visual memories. For each candidate window, the analyzer builds a compact listener packet:

$$P_i = B_\phi\left(T_i, E_i^{local}, M_{i-1}^{scene}, R_i, M_{i-1}^{id}, M_{i-1}^{vis}, Q_i\right).$$

Here E_i^{local} contains local interaction evidence derived from the aligned trace and audio analysis, R_i contains retrieved memory atoms and callback candidates, and Q_i contains pacing information, uncertainty notes, and provenance metadata. The packet P_i is the bounded input inspected by a dedicated listener module. The listener does not directly consume the raw audio, full transcript, full memory store, or unbounded campaign history.

The multi-stage listener applies four sequential calls to this same packet. Let L_ω be an evaluation function of moment relevance. The listener builds three evidences:

$$e_i^{local} = L_\omega^{local}\left(P_i, \emptyset\right).$$

$$e_i^{scene} = L_\omega^{scene}\left(P_i, \{e_i^{local}\}\right).$$

$$e_i^{continuity} = L_\omega^{continuity}\left(P_i, \{e_i^{local}, e_i^{scene}\}\right).$$

These tasks denote different task prompts applied to the same compact packet. The local prompt emphasizes passage-level support, the scene prompt emphasizes scene-state changes and immediate consequences, and the continuity prompt emphasizes retrieved atoms, callbacks, active threads, prior choices, unresolved threats, and recent visual memories. The decision, and the listener call, are defined as:

$$d_i = L_\omega^{final}\left(P_i, \{e_i^{local}, e_i^{scene}, e_i^{continuity}\}\right).$$

The decision record d_i contains the visualization decision, confidence, rationale, evidence used, counterevidence, uncertainties, visualization strategy, grounded elements, placeholder elements, forbidden or unknown details, and director-facing continuity notes. The final call does not write the image direction prompt. A binary commitment variable is obtained after normalization and grounding validation:

$$z_i = \mathbf{1}\left[d_i.\text{decision} = \text{visualize} \ \wedge \ G(d_i, P_i) = \text{valid}\right]$$

where G denotes an LLM-based grounding and normalization step applied before downstream use. In practice, G checks whether the decision conforms to the expected schema, whether its claims are supported by evidence in P_i, and whether the proposed visualization contains concrete scene elements grounded in the packet rather than unsupported details. Decisions that fail these checks are treated as non-visual and are not committed. For each committed window, the record is:

$$v_i = \big\langle W_i,\ T_i,\ P_i,\ e_i^{local},\ e_i^{scene},\ e_i^{continuity},\ d_i,\ D_\nu(P_i, d_i),\ I_i \big\rangle.$$

Here $D_\nu(P_i, d_i)$ is the image direction prompt produced by the director module, and I_i is the rendered image. The image materializes a selected and grounded record rather than serving as the objective optimized by the architecture.

Hierarchical Multimodal Memory: The memory layer constructs a common temporal representation for heterogeneous evidence. Actual-play salience is not expressed only through transcript semantics. It may appear in timing, overlap, silence, laughter, short reactions, speaker uncertainty, or callbacks to earlier events. We align transcript, speaker, timing, and interaction evidence to a shared decision-window representation, with turn-level annotations when the signal supports that granularity. ASR provides timestamped transcript segments, and diarization provides speaker intervals. For each decision window, transcript segments are clipped to the window boundaries and aligned with speakers.

Additional evidence is added through signal processing, deterministic aggregation, and pretrained audio models. Voice activity is estimated with Silero VAD (Team 2024b). Speaker diarization and overlapped-speech evidence are derived with pyannote audio (Bredin et al. 2020). Audio-event cues are extracted with YAMNet, an AudioSet-trained event classifier (Gemmeke et al. 2017). Prosodic summaries are extracted from the waveform with librosa, and pitch-related cues are computed with Parselmouth/Praat (McFee et al. 2015; Jadoul, Thompson, and De Boer 2018). Dimensional speech-emotion estimates are produced with the audeering wav2vec2 MSP-dimensional emotion model (Wagner et al. 2023). These outputs are aggregated into evidence records

describing turn structure, speaker switches, overlap duration, gap statistics, rapid exchange, long post-turn silence, short-reaction clusters, crosstalk, nonverbal audio events, and changes in arousal and valence.

Longer-range context is maintained through scene state, memory atoms, and an active memory bundle. This treatment of memory connects to situation-model accounts of narrative comprehension. Zwaan and Radvansky (1998) describe comprehension as constructing an integrated representation of the described state of affairs. This motivates our use of the many memory modules as compact traces of what is currently true, what has changed, and why a moment may matter. At the start of an evaluation, the system loads any atoms saved from previous game sessions and inserts them into a memory embedding index. During analysis, new atoms are created when maintenance detects a narrative update that should persist beyond the current window. These updates include discoveries, decisions, threats, promises, unresolved questions, changes in state, and other information likely to affect later interpretation. A committed visual memory is also converted into an atom, so selected moments become retrievable context for future decisions.

Each stored atom contains an identifier, timestamp, type, claim, open label, evidence, confidence, status, source excerpt, source window id, and normalized tags for entities, threads, objects, locations, and motifs. Fresh atoms are appended to the session atom list and immediately added to the embedding index, using the claim and source excerpt to build their textual representation. Search also uses atom metadata, including entities, threads, objects, locations, and motifs, for heuristic retrieval. The final analysis output stores the updated memory state and the full atom list, allowing later packets and later sessions to retrieve the accumulated narrative memory and to evaluate continuity.

Multi-stage Listener Reasoning: The listener module turns accumulated evidence into a selection judgment. It does not operate directly on the raw transcript, audio features, or memory store. Instead, each candidate packet presents a bounded view of the current interval, including the transcript excerpt, selected turns, local interaction evidence, scene state, retrieved atoms, identity context, recent visual memories, pacing information, and uncertainty notes. The listener is asked whether this interval has enough local support, narrative consequence, and visual grounding.

The decision process uses four sequential listener calls. The first three produce evidence records rather than final decisions, each containing grounded claims, supporting evidence, confidence estimates, and uncertainties. Later calls receive the same candidate packet plus earlier records, allowing evidence to accumulate before the final judgment. The local-evidence call checks whether the current passage contains quoted turns, speaker actions, reactions, acoustic or conversational cues, affective shifts, and visualizable elements. The scene-relevance call assesses whether the passage changes, clarifies, or intensifies the current fictional situation, including goals, conflict, stakes, participants, boundaries, and immediate consequences. The continuity-relevance call evaluates longer-range connections through retrieved atoms, callbacks, active threads, prior choices, unresolved threats, recurring motifs, prior visual memories, and continuity constraints.

The final listener call uses the current passage as the primary source of truth, while using memory, identity context, recent visual memories, and the most recent generated prompt to check for inconsistency or repetition. A positive decision does not produce the image direction prompt; instead, it records grounded elements, placeholders, unknown or forbidden details, and director-facing constraints. Before downstream use, the output is parsed, normalized, and validated, and positive decisions without packet-grounded evidence or concrete visual anchors are treated as non-visual. This separates the listener's responsibility for deciding whether a moment should be externalized from the director's responsibility for composing the image.

Visual Externalization and Grounding: A director module compiles the validated listener decision into an image direction prompt. It receives the compact packet and the validated decision, including moment description, supporting evidence, visualization strategy, grounded elements, forbidden or unknown details, and editorial notes. The director's role is to stage the approved material as a single prompt with a clear visual core, setting, composition, character placement, objects, and style. Rendering is treated as a downstream step. The system submits validated prompt packages to an image-generation backend and stores each rendered image with its source interval, packet, listener decision, and director output. The resulting visual artifact remains traceable to the evidence and decision that produced it, rather than serving as part of the selection mechanism.

Experiments

The evaluation uses two actual-play sessions in complementary roles: a public English-language session for the quantitative moment-selection evaluation, hereinafter E1, and a Portuguese-language homebrew session for the perceptual study of the generated images, hereinafter E2. Experiments were performed on an NVIDIA H200 (80 GB HBM3). Automatic speech recognition used WhisperX (Bain et al. 2023) with the large-v3 checkpoint (Radford et al. 2023) (bfloat16, batch size 16, beam size 1); speaker diarization used pyannote/speaker-diarization-3.1 (Bredin et al. 2020). Memory-atom retrieval used all-MiniLM-L6-v2 sentence embeddings (Reimers and Gurevych 2019). The listener and director stages were served by Qwen2.5-32B-Instruct (Team 2024a) through a locally hosted vLLM server (Kwon et al. 2023) (8,192-token context, 256 max concurrent sequences). Image rendering used FLUX.2-dev (Labs 2025) through a ComfyUI backend at 1024×1024 resolution: a 25-step first pass (guidance scale 1.25) followed by an 8-step refinement pass (guidance scale 1.0, denoise strength 0.2). End-to-end wall-clock time was dominated by transcription and diarization: 505 s and 309 s, respectively, for the 3h 25m episode, and 108 s and 33 s for the 55-minute session.

Public actual-play silver label protocol (E1): We evaluate moment selection on a public long-form actual-play session (video id `byva0hOj8CU`), of duration $L = 12{,}271$ s

($\approx$ 3h 25m). The preprocessing stage produced 205 one-minute decision windows, of which the listener committed 121 as visual memories, a coverage of $\approx 59.0\%$. Following prior work that treats audience reactions as a proxy for highlights (Fu et al. 2017), we derived silver labels for viewer-perceived salience from two independent signals, shown over the full episode in Figure 3. Briefly, a silver label is an annotation obtained indirectly from observable data rather than through manual ground-truth labeling. Such labels are useful when direct annotation is unavailable or costly, but they should be interpreted as noisy proxies for the underlying construct rather than as definitive ground truth. In our case, the underlying construct is whether a moment was salient to viewers, while the observable signals are the timestamped comments and replay behavior. The bottom plot shows the timestamps manually typed by viewers in the video's public comments. Of 5,373 retrievable comments, 300 distinct timestamps were extracted from 1,414 total mentions. The top image is the video's "most-replayed" engagement heatmap, a signal previously used to derive highlight labels at scale (Sul, Han, and Lee 2023), from which we took the top-20% most-replayed segments.

At temporal tolerances $\tau \in \{0, 30, 60\}$ s, we report the recall of silver-label timestamps (a timestamp is recalled if a committed window lies within τ), the frequency-weighted recall (each timestamp weighted by its citation count), and the window-level precision (the fraction of committed windows containing at least one silver-label timestamp within τ). Because coverage is high, raw recall is not informative on its own: a selector that commits 59% of windows recalls many timestamps by construction. We therefore compare each observed value against a chance model of 2,000 random selections of the same window count (seed 0), and report the mean baseline recall, the lift (observed / baseline), and the distance of the observed value from the baseline mean in standard deviations (z).

Human evaluation protocol (E2): We perform an image study using part of a Portuguese-language homebrew session (a home Dungeons & Dragons campaign recorded over digital call with multiple players), of duration $\approx$ 55 minutes, split into 55 one-minute decision windows. The listener committed 28 of the 55 windows as visual memories, a coverage of 50.9%. The players who participated in this session were asked to evaluate both the generated image content and the notoriety of the selected moments. To assess the generated images directly, we conducted a perceptual study over a seeded random sample of 10 of the 28 committed visual memories from the Portuguese-language homebrew session. The five volunteer raters were players from the recorded campaign. All were known to the authors, but none is an author of this paper. Each had more than ten years of experience with tabletop role-playing games and, at the time of recording, had participated in this campaign for approximately two years, playing roughly once per month. The raters were not informed of the objective of the study or of the details of the proposed approach before evaluation.

Using campaign participants as raters presents a deliberate trade-off. Because they participated in the fiction, they can judge whether an image is grounded in what actually occurred at the table, including contextual details that an external evaluator would not be able to verify from the evaluation material alone, and can therefore serve as domain-expert evaluators. At the same time, they are not independent of the material, may rely on campaign knowledge not contained in the one-minute window, and may be invested in the campaign in ways that bias their judgments. We therefore treat the study as descriptive rather than generalizable evidence, with samples drawn from a fixed seed.

τ (s)	Recall	Wgt. Rec.	Prec.	F1	Base.	Lift
0	71.0%	72.5%	73.6%	0.723	61.0%	1.16×
30	88.3%	91.0%	89.3%	0.888	83.8%	1.05×
60	94.7%	96.3%	94.2%	0.944	93.6%	1.01×

Table 1: Moment-selection performance against 300 viewer-referenced timestamps: recall, frequency-weighted recall, window-level precision, F1, chance baseline (2,000 random samples), and lift over chance. The distance from chance (z) is 2.3, 1.2, and 0.4 at $\tau = 0, 30, 60$ s.

Results

Table 1 reports moment-selection performance against the 300 comment-derived silver labels on E1, and Figure 4 shows the committed windows and silver-label timestamps over the full episode. At $\tau = 0$ (the exact decision window), the listener recalled 71.0% of viewer-referenced timestamps (213/300) against a chance baseline of $61.0\% \pm 4.4$, a lift of $1.16\times$ ($z = 2.3$); window-level precision was 73.6% (89/121 committed windows contained a silver-label timestamp). Frequency-weighted recall was slightly higher, at 72.5%. As the tolerance widened, observed recall rose to 88.3% ($\tau = 30$ s) and 94.7% ($\tau = 60$ s), with window-level precision rising to 89.3% (108/121) and 94.2% (114/121), but the chance baseline rose as well (83.8% and 93.6%), so the lift contracted to $1.05\times$ ($z = 1.2$) and $1.01\times$ ($z = 0.4$).

Figure 5 shows the observed recall converging to the baseline as the tolerance grows. Among the 87 timestamps not covered at $\tau = 0$, the median gap to the nearest committed window was 27 s (mean 41 s, maximum 191 s), so most misses were near-misses rather than distant ones. Against the second salience signal, committed windows covered 60.0% (12/20) of the top-20% most-replayed heatmap segments; this value is indistinguishable from the equal-coverage chance baseline ($59.8\% \pm 10.3$ under the same 2,000-selection procedure; lift $1.00\times$, $z = 0.0$), so the heatmap signal provides no evidence of targeted selection. With only 20 coarse segments over a 3.4-hour episode, however, this test also has little statistical power, and we report it for completeness rather than as a negative finding.

The upstream candidate set — the 133 evidence windows flagged before final listener commitment, a coverage of 64.9% — yields 74.7% recall and 71.4% precision at $\tau = 0$ (F1 0.730), rising to 92.0%/87.2% at $\tau = 30$ s and 96.7%/93.2% at $\tau = 60$ s. The final commitment step therefore acts as a mild precision filter: relative to the candidate set, it lowers coverage by 5.9 points and exact-window recall

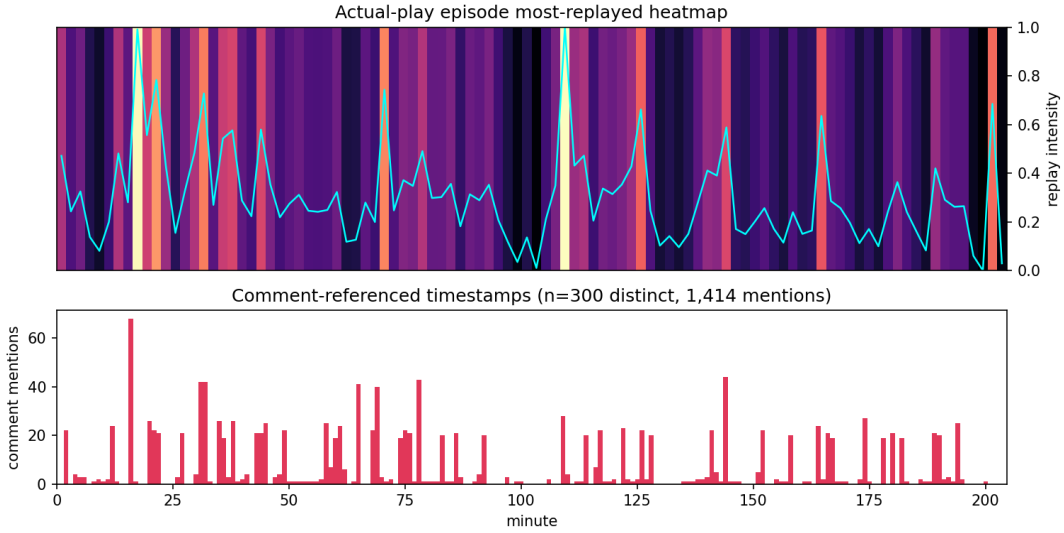


Figure 3: Silver-labels of the public actual-play episode. Top: the video's "most-replayed" engagement heatmap. Bottom: viewer-typed timestamps aggregated per minute and weighted by citation frequency (300 distinct timestamps, 1,414 mentions).

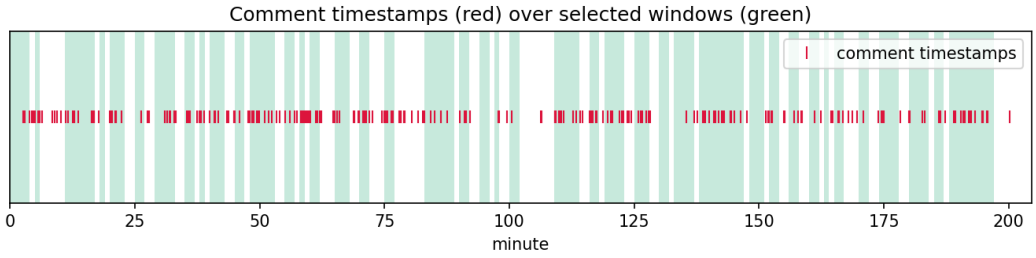


Figure 4: Committed decision windows (green bands) and viewer-referenced comment timestamps (red ticks) over the public actual-play episode. The listener committed 121 of 205 one-minute windows (59.0% coverage); at exact-window resolution the committed set contains 71.0% of the 300 silver-label timestamps.

by 3.7 points while raising precision by 2.2 points.

The results indicate that the listener aligns with viewer-perceived salience above chance only at exact-window resolution. At $\tau \geq 30$ s the high absolute recall reflects the 59.0% window coverage rather than targeted selection: the observed value then falls within one-and-a-half ($\tau = 30$ s) and half ($\tau = 60$ s) baseline standard deviations of chance. The consistently higher frequency-weighted recall suggests that the selector lands on more frequently cited moments somewhat more often than on rarely cited ones. We read this as a mild alignment with viewer attention rather than evidence of a mechanism, since the evaluation is observational: comment frequency and selection may both track underlying scene salience without a direct link between them. The claim these results support is accordingly narrow: the memory-based listener selects more viewer-referenced windows than an equal-coverage random baseline at exact-window resolution. They are not evidence that the selected images are faithful, useful during play, or preferred by players; those questions are the object of the human evaluation.

The human evaluation on E2 was performed in two rounds. In a first round, raters listened to a one-minute audio segment and were shown the corresponding generated image. They then answered a five-point Likert item, "*How well does the image match this audio moment?*" (1 = "*not at all*", 5 = "*perfectly*"), with an optional free-text comment. Every rater judged all ten items, yielding 50 ratings in total. In a second round, we conducted a pairwise preference study with the same five raters comparing images generated from director-authored prompts with 11 images generated directly from the raw transcript, yielding 55 pairwise judgments in total. For each scene, raters were shown a short textual context derived from the session transcription, consisting of a brief paragraph summarizing the events immediately preceding the scene and a two-sentence summary of the scene itself. This context was presented alongside the two images, and raters were asked which they preferred, with a "neither/no preference" option. They were not informed which generation procedure had produced either image.

Figure 6 shows one rated item from the perceptual study (E2): the one-minute transcript excerpt for a committed window, the director prompt generated from it, the resulting image as presented to raters in the form. For comparison, the Figure also shows a separate image generated directly from the raw transcript for the same scene; the pairwise study quantifies how raters judge this comparison across scenes.

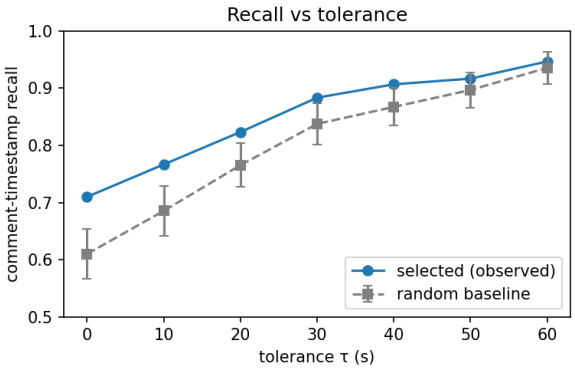


Figure 5: Silver-label recall of the committed windows against the equal-coverage random baseline (mean $\pm$ one standard deviation over 2,000 selections) as the temporal tolerance τ grows. The gap over chance is concentrated at exact-window resolution and closes as the tolerance widens.

Likert score	% of ratings
5 (perfect match)	26.0
4	26.0
3	22.0
2	22.0
1 (no match)	4.0
Mean (SD)	3.48 (1.20)

Table 2: Human ratings of image–audio match for a seeded sample of 10 visual memories, each judged by five raters.

Across the 10 rated visual memories evaluated in E2 and judged by five raters (50 ratings), the mean image–audio match score was 3.48 on the five-point scale (SD 1.20; median 4), and 52.0% of ratings were 4 or 5. Inter-rater agreement was modest (Krippendorff's ordinal $\alpha = 0.34$), so per-item means should be read as noisy estimates; we therefore complement them with the pooled distribution and the raters' free-text comments. Table 2 summarizes the score distribution, and Figure 7 shows the individual ratings per item. Item means ranged widely, from 2.4 to 4.6: the two highest-rated items (mean 4.6) were single-narrator descriptive scenes (an armored guard entering a tavern; a character concentrating on a telepathy spell), while the two lowest (mean 2.4) failed for different reasons discussed below.

The free-text comments point to three recurring failure modes. The first is a moment-selection failure: one audio window was identified by raters as containing no concrete scene to illustrate ("it didn't have anything to work with"), and the resulting keyframe was rated 1 or 2 by three of the five raters. This indicates that the listener can still commit windows without enough visualizable substance, and that relevance filtering before externalization can be improved. The second is partial grounding: images that stage the right scene but omit described participants or the central action; the item rated 2 by four of the five raters showed "only half of the scene", dropping the interaction the audio narrates, and even the one milder rating on that item cited a missing participant. The third is cast and identity drift: several comments note more figures than active speakers, near-duplicate faces across distinct characters, and ungrounded embellishments (a drawn sword, a fire arrow) that the audio never describes. On the positive side, the highest-rated items were praised for capturing staging and mood ("gives a good sense of gravity to the attempt") suggesting that when the committed window contains a concrete, narratable event, the director stages it into a scene raters recognize. Consistent with the moment-selection evaluation, these observations support a narrow claim: the pipeline produces recognizably grounded images for well-chosen moments, while the main quality bottlenecks are committing low-content windows.

In a second round, the same five raters directly compared the pipeline's images against a simpler alternative: an image generated straight from the raw one-minute transcript, without narrative memory, scene context, or director-stage composition. Across the 55 pairwise judgments, raters preferred the pipeline's image in 40 cases (72.7%), the raw-transcript image in 4 (7.3%), and expressed no preference in the remaining 11 (20.0%). This shows the pipeline's images being preferred in 90.9% of cases in which a preference was expressed ($p \approx 2 \times 10^{-8}$). Further, in 4 scenes the choice of the pipeline-generated image was unanimous, while the raw-transcript image was never preferred by the majority in any scene. Raters chose "neither" as the majority response in the two scenes whose comments also described nothing concrete to depict, the same moment-selection failure noted in E1. One scene split preference evenly between the two images, as it consisted of a continuous, single-narrator passage and worked similarly well for both the raw-transcript and the director prompt. The raw-transcript image's losses otherwise trace to the same failure modes raised qualitatively above. For instance, one comment on the raw-transcript image for a scene built around a player's in-character exclamation of "you fiend" noted that "[the depiction of a demon] has nothing to do with the scene", a literal reading that was avoided by the pipeline's memory-grounded prompt. Other comments on losing raw-transcript images cite garbled on-image text, extraneous or duplicated figures, and compositions that "make little sense". Taken together, this pairwise controlled test implies that raters prefer our images to a raw-transcript baseline by a wide and significant margin, justifying such an approach in comparison to the naive baseline.

Limitations: The evaluation covers one public actual-play session for moment selection and one separately recorded home campaign session for image–audio match, so the findings remain scoped to the evaluated systems, table cultures, and languages. Both sessions nevertheless use the same fully traced pipeline. Moment selection evaluation relies on viewer-typed timestamps and replay activity, which are noisy and temporally imprecise proxies for viewer salience rather than direct measures of whether a passage supports a grounded image. Because the system commits 59% of windows, raw recall is less informative on its own, so we report lift over an equal-coverage chance baseline; the observed advantage is concentrated at $\tau = 0$ s. Image–audio alignment

Raw transcript **Director's prompt**



Transcript excerpt (1-minute window; translated from Portuguese): *"[SPEAKER_02] How much damage is that, roughly? [SPEAKER_00] It's this here, look. [SPEAKER_02] Damn, is it the magic shot or the normal shot? [SPEAKER_00] It's the magic shot. [SPEAKER_02] Ok. [SPEAKER_02] This hit is strong enough to hit the... [SPEAKER_00] It's fire. [SPEAKER_02] Ok. [SPEAKER_00] And... oh wait, no, it's actually bludgeoning. [SPEAKER_02] I think even if that's not enough for the bear, it feels quite hurt and realizes that... [SPEAKER_02] It's not doing okay [SPEAKER_02] Ok. [SPEAKER_02] Right, Leon, you now see the bear has fled and you're... You see a bright arrow hitting the bear. [SPEAKER_02] You're face to face with this man... [SPEAKER_00] Who said I was face to face?"*

Director prompt: *"Digital painting, RPG concept art, an intricate fantasy illustration. A magical fire arrow streaks through the air, its fiery trail illuminating the dense forest. It strikes a massive bear, causing it to roar in pain and fury. The bear's fur smolders, and its eyes widen with shock and fear. The forest around them is dark, with dappled moonlight filtering through the canopy, casting eerie shadows. The scene is tense, with a sense of danger and the aftermath of a powerful magical attack hanging in the air."*

Figure 6: A committed visual memory from the perceptual study, from a window with two active speakers, mean rating $4.0/5$ across the five raters. The transcript records the table correcting the attack's damage type from fire to bludgeoning mid-exchange, but the director's prompt and rendered image still stage it as a fire arrow.



Figure 7: Image–audio match ratings per item. Dots are individual ratings while bars show means. Low spread indicates strong inter-rater agreement, and the lowest ratings are for items 03 (partial rendering) and 09 (limited visual context).

is assessed separately in the perceptual study. The perceptual study is small and uses campaign participants as raters, whose familiarity supports informed grounding judgments but may also introduce bias or reliance on context beyond the presented window. It evaluates the pipeline largely as a whole, although free-text comments reveal recurring patterns and the raw-transcript and upstream-candidate comparisons isolate prompt grounding and final commitment.

Conclusion

This work focused on answering two questions: whether the significant moments of a long-form actual-play TTRPG session can be identified automatically from its audio, and whether those moments can be visualized as memory-evoking images. Our approach couples narrative memory to selection: heterogeneous audio evidence is aligned to short decision windows, and a multi-stage LLM listener judges each against the current scene and the session's accumulated memory before committing it as a *visual memory*, a structured record traceable from evidence to decision to image.

Our experiments support a deliberately narrow positive answer to both questions. On a publicly available actual-play episode, the listener recovered viewer-referenced timestamps above chance at exact-window resolution (71.0% recall against $61.0\% \pm 4.4$, a lift of $1.16\times$), an alignment with viewer-perceived salience that fades as the temporal tolerance grows, and that the coarser most-replayed signal is not sufficiently relevant to confirm. On a Portuguese-language homebrew session, players of the recorded campaign rated the association between generated images and their source passages at 3.48 of 5 on average, with just over half of the ratings at 4 or 5. The highest-rated items were concrete, narratable scenes, and the recurring failures (committed windows with little to visualize, partial renderings of a scene, and ungrounded artifacts) point to moment selection and grounding enforcement, rather than rendering quality, as the main bottlenecks. The pairwise round showed raters strongly preferring the pipeline's images over a raw-transcript baseline. These results also show the potential for non-professional campaigns, with different linguistic and cultural contexts or less sophisticated recording setups.

These findings position visual externalization in actual play as a selection problem that narrative memory makes tractable, but that remains unsolved. The pairwise comparison confirms that simply committing raw transcripts is ineffective in properly conveying grounded visual memories. The proposed approach significantly improves the perceived quality of this naive baseline, both in finding memorable moments and in externalizing them. Beyond the component contrasts outlined in the Limitations, we see the following directions as most promising: First, experimenting with alternative downstream tasks, bringing other ways to enhance the game experience, whether with immersion or utility; Second, comparing the results obtained in our experiments against alternative baselines, as well as against the same pipeline using different models; And, ultimately, a live study of how memory-supported visual externalization affects player engagement and immersion at the table.

Acknowledgments

This work was supported by the authors' individual grants from the Kunumi Institute.

References

Acharya, D.; Mateas, M.; and Wardrip-Fruin, N. 2021. Story improvisation in tabletop roleplaying games: Towards a computational assistant for game masters. In *2021 IEEE Conference on Games (CoG)*, 01–08. IEEE.

Bain, M.; Huh, J.; Han, T.; and Zisserman, A. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *INTERSPEECH 2023*.

Best, T.; Lucas, S.; and Gaina, R. 2024. 'Journeys in the dark' - towards game master AI in complex board games. In *Proceedings of the Twentieth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE '24. AAAI Press. ISBN 1-57735-895-3.

Borecky, A. 2021. Dungeons, Dragons, and Music: The Immersive Qualities of Sound in Dungeons & Dragons. *Journal of Sound and Music in Games*, 2(1): 46–64.

Bredin, H.; Yin, R.; Coria, J. M.; Gelly, G.; Korshunov, P.; Lavechin, M.; Fustes, D.; Titeux, H.; Bouaziz, W.; and Gill, M.-P. 2020. Pyannote. audio: neural building blocks for speaker diarization. In *ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 7124–7128. IEEE.

Brown, E.; and Cairns, P. 2004. A grounded investigation of game immersion. In *CHI '04 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '04, 1297–1300. New York, NY, USA: Association for Computing Machinery. ISBN 978-1-58113-703-3.

Busselle, R.; and Bilandzic, H. 2009. Measuring narrative engagement. *Media psychology*, 12(4): 321–347.

Callison-Burch, C.; Tomar, G. S.; Martin, L. J.; Ippolito, D.; Bailis, S.; and Reitter, D. 2022. Dungeons and Dragons as a Dialog Challenge for Artificial Intelligence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 9379–9393. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Cover, J. G. 2014. *The creation of narrative in tabletop role-playing games*. McFarland.

Deterding, S.; and Zagal, J. 2018. *Role-playing game studies: Transmedia foundations*. Routledge.

Ellis, S.; and Hendler, J. 2017. Computers Play Chess, Computers Play Go. . . Humans Play Dungeons & Dragons. *IEEE Intelligent Systems*, 32(4): 31–34.

Ferreira, L. N.; Lelis, L. H. S.; and Whitehead, J. 2020. Computer-generated music for tabletop role-playing games. In *Proceedings of the Sixteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, AIIDE'20, 59–65. AAAI Press. ISBN 978-1-57735-849-7.

Fine, G. A. 2002. *Shared fantasy: Role playing games as social worlds*. University of Chicago Press.

Fu, C.-Y.; Lee, J.; Bansal, M.; and Berg, A. C. 2017. Video Highlight Prediction Using Audience Chat Reactions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 972–978. Association for Computational Linguistics.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 776–780. IEEE.

Hanjalic, A.; and Xu, L.-Q. 2005. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7: 143–154.

Jadoul, Y.; Thompson, B.; and De Boer, B. 2018. Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71: 1–15.

Jennett, C.; Cox, A. L.; Cairns, P.; Dhoparee, S.; Epps, A.; Tijs, T.; and Walton, A. 2008. Measuring and defining the experience of immersion in games. *International Journal of Human-Computer Studies*, 66(9): 641–661.

Kellogg, R. T.; Whiteford, A. P.; Turner, C. E.; Cahill, M.; and Mertens, A. 2013. Working memory in written composition: A progress report. *Journal of Writing Research*, 5(2): 159–190.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, 611–626. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702297.

Labs, B. F. 2025. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2.

Lei, J.; Berg, T. L.; and Bansal, M. 2021. QVHighlights: Detecting Moments and Highlights in Videos via Natural Language Queries. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*.

Li, Y.; Gan, Z.; Shen, Y.; Liu, J.; Cheng, Y.; Wu, Y.; Carin, L.; Carlson, D.; and Gao, J. 2019. StoryGAN: A Sequential Conditional GAN for Story Visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6329–6338.

Ma, Z.; Zheng, Z.; Ye, J.; Li, J.; Gao, Z.; Zhang, S.; and Chen, X. 2024. emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 15747–15760. Bangkok, Thailand: Association for Computational Linguistics.

Maharana, A.; Hannan, D.; and Bansal, M. 2022. StoryDALL-E: Adapting Pretrained Text-to-Image Transformers for Story Continuation. In *Computer Vision – ECCV 2022*, Lecture Notes in Computer Science. Springer.

McFee, B.; Raffel, C.; Liang, D.; Ellis, D. P.; McVicar, M.; Battenberg, E.; Nieto, O.; et al. 2015. librosa: Audio and music signal analysis in python. *SciPy*, 2015(18-24): 7.

Padovani, R.; Ferreira, L.; and Lelis, L. 2017. Bardo: Emotion-Based Music Recommendation for Tabletop Role-Playing Games. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 13(1): 214–220.

Paivio, A. 1990. *Mental representations: A dual coding approach*. Oxford university press.

Petersen, S.; Willis, L.; Mason, M.; and Fricker, P. 2016. *Call of Cthulhu: Keeper Rulebook*. Ann Arbor, MI: Chaosium Inc., 7th edition. ISBN 978-1-56882-430-7.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision.

Rameshkumar, R.; and Bailey, P. 2020. Storytelling with dialogue: A critical role dungeons and dragons dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5121–5134.

Reagan, A. J.; Mitchell, L.; Kiley, D.; Danforth, C. M.; and Dodds, P. S. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1): 31.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. Hong Kong, China: Association for Computational Linguistics.

Russell, J. A. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6): 1161–1178.

Shyne, F.; and Cooper, S. 2024. No Player Left Behind: Evolving Dungeons and Dragons Combat to Optimize Difficulty and Player Contributions. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, 20(1): 224–234.

Sul, J.; Han, J.; and Lee, J. 2023. Mr. HiSum: A Large-scale Dataset for Video Highlight Detection and Summarization. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Track on Datasets and Benchmarks*.

Sun, M.; Farhadi, A.; and Seitz, S. 2014. Ranking Domain-Specific Highlights by Analyzing Edited Videos. In *Computer Vision – ECCV 2014*, volume 8689 of *Lecture Notes in Computer Science*, 787–802. Springer.

Team, Q. 2024a. Qwen2.5 Technical Report. arXiv:2412.15115.

Team, S. 2024b. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad.

Too, S.-W.; See, J.; Quek, A.; and Goh, H.-N. 2023. EmoStory: Emotion Prediction and Mapping in Narrative Stories. *JOIV : International Journal on Informatics Visualization*, 7(3-2): 2048–2055.

Wagner, J.; Triantafyllopoulos, A.; Wierstorf, H.; Schmitt, M.; Burkhardt, F.; Eyben, F.; and Schuller, B. W. 2023. Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10745–10759.

Wang, S.; and Ji, Q. 2015. Video Affective Content Analysis: A Survey of State-of-the-Art Methods. *IEEE Trans. Affect. Comput.*, 6(4): 410–430.

Wizards of the Coast. 2014. D&D Beyond, Basic Rules (2014). Available at: https://www.dndbeyond.com/sources/dnd/basic-rules-2014/introduction. Accessed July 1, 2026.

Zhao, Z.; Zhu, X.; Wang, X.; Wang, S.; Geng, X.; Tian, W.; and Xie, L. 2025. Steering language model to stable speech emotion recognition via contextual perception and chain of thought. *IEEE Transactions on Audio, Speech and Language Processing*, 34: 415–426.

Zhu, A.; Aggarwal, K.; Feng, A.; Martin, L. J.; and Callison-Burch, C. 2023a. FIREBALL: A Dataset of Dungeons and Dragons Actual-Play with Structured Game State Information. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4171–4193. Toronto, Canada: Association for Computational Linguistics.

Zhu, A.; Martin, L.; Head, A.; and Callison-Burch, C. 2023b. CALYPSO: LLMs as dungeon masters' assistants. In *Proceedings of the Nineteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 19 of *AIIDE '23*, 380–390. Salt Lake City: AAAI Press. ISBN 978-1-57735-883-1.

Zwaan, R. A.; and Radvansky, G. A. 1998. Situation models in language comprehension and memory. *Psychological bulletin*, 123(2): 162.