



StoryFX: Augmenting Long-form Audio Stories with Generated Sound Effects Through Large Language Models and Text-to-Audio Generators

Thant Zaw^a, Thinn Thinn Wai^b

^aFaculty of Computer Science, University of Information Technology, Yangon, Myanmar

^bFaculty of Computer Science, University of Information Technology, Yangon, Myanmar

DOI : <https://doi.org/10.55248/gengpi.07.1026.3275>

ABSTRACT :

Contemporary long-form audio stories in audiobook and podcast formats are limited to speech-only methods for storytelling. Integrating sound effects (SFX) through discrete event sounds and continuous ambient soundscapes can enhance listener engagement, emotional impact, and perceived presence, but manual sound design is labor-intensive and costly. This paper presents StoryFX, an end-to-end framework for automatically translating narrative text into synchronized sound-effect compositions using large language models (LLMs) for semantic orchestration, text-to-audio (TTA) models for sound synthesis, and digital audio workstations (DAWs) for audio composition. StoryFX introduces a temporal-semantic ontology for decomposing audiobook sound design into interpretable subtasks, enabling task-specific prompting and stepwise evaluation, together with a collection of multi-strategy prompting templates for inference optimization. Qwen 2.5 32B Instruct and Woosh were employed as foundation models to adapt a human-annotated story excerpt. We evaluate the framework using Qwen2.5-32B-Instruct for orchestration and Woosh for sound generation on a human-annotated audiobook excerpt.

Keywords:Text-to-Audio Generation, Large Language Models, Prompt Engineering, Audio Systems, Audio Storytelling, Immersive Audio, Audiobooks, Podcasting, Sound Studies

Introduction

Due to multi-tasking needs in the contemporary lifestyle, auditory media have increased in popularity, facilitated by major developments in streaming services and mobile technology. In addition to music, long-form narrative formats such as audiobooks and storytelling podcasts have become staple choices for listeners [1]–[3]. Scholarly inquiry in this domain has focused mainly on the educational or business perspectives of these formats, despite reports revealing entertainment as the primary motive for listening [2], [4]. Consequently, research should develop toward the examination and innovation of entertainment mechanisms in narrative sound design.

Although sound can serve as a rich storytelling medium, the creative methodology of contemporary audio stories is largely confined to the spoken-word model. Such an audio creation model relies solely on the vocal delivery of the narrators and characterization of the voice actors for dramatic and immersive rendition of the story [5]. Compared to the multi-modal complexity of sound design in audiovisual media such as films, animations, and video games, the voice-only approach offers a narrower range of aesthetic representation.

To that end, the listening experience of audio stories can be expanded by incorporating sound effects (SFX) alongside verbal narration to depict and dramatize the occurrences in the story. Specifically, these are diegetic SFX that represent the (i) foreground plot events such as character actions and non-verbal vocalizations (NVV), and sudden material movements etc., and (ii) background scene ambience such as weather phenomena, environmental noises, room tone etc.,

The historical precedent for a SFX-integrated audio narrative format is provided by the mainstream success of radio drama and radio features in the 20th century before the mass adoption of television. In radio drama production, live performances of a voice cast, musicians, and Foley artists were combined to form a multi-modal audio film [6], [7]. Beyond aesthetic novelty, SFX can also have structural influences on the audience. For example, studies have shown that the introduction of meaningful non-verbal sounds to listeners can result in attentional elicitation, emotional evocation, and an improved sense of presence [8]–[11].

However, manual adaptation of long-form audio stories into SFX-augmented soundscapes would require more labor, time, and expertise in-scalable with longer works that can span up to tens of hours. The production process would involve additional hours of analyzing the text material, formulating a SFX cue-sheet, sourcing or composing the assets, and mixing the SFX with the narrator's voice.

To address this issue, this paper introduces StoryFX, an automatic end-to-end SFX composition framework using a Large Language Model (LLM) for semantic and discourse analysis of the long-form text, a Text-to-Audio Model (TTA) for the synthesis of raw SFX assets, and a Digital Audio

Workstation (DAW) for adding spatial effects and seamless synchronization with the vocal stream. The LLM component interprets the story event and scene indicators from the complex descriptions into contextually informed SFX cues, while the TTA and mixing components create a narratively aligned SFX composition based on these cues.

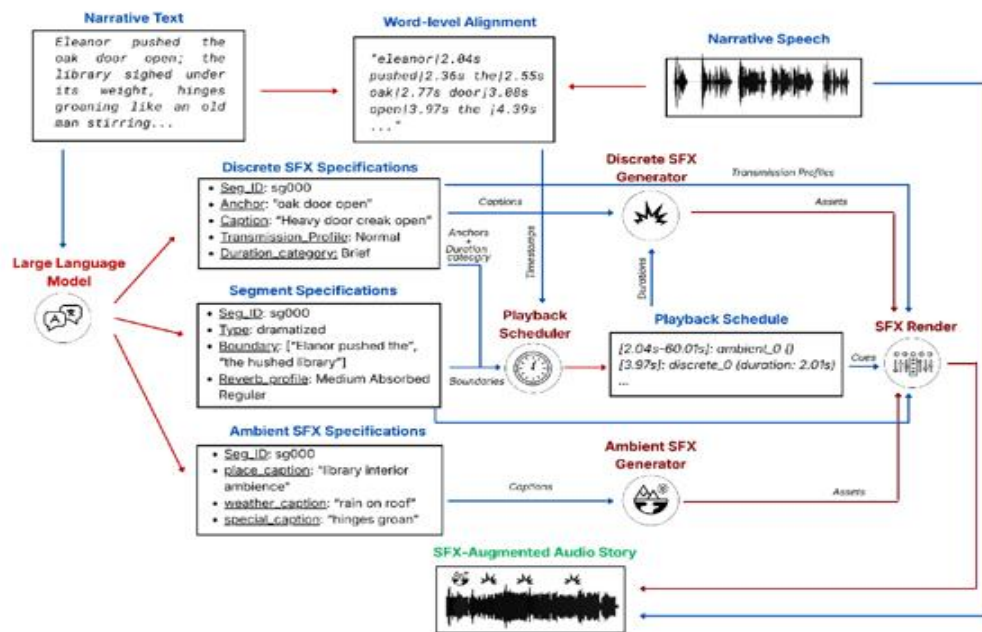


Fig.1 – System Overview of StoryFX

Related Work

The design and feasibility of StoryFX is based on the recent advances in foundational LLMs and TTAs, as well as automated narrative SFX composition frameworks.

1.1. Large Language Models (LLM)

LLMs are transformer-based neural language models trained on large-scale natural language corpora to perform probabilistic next-token prediction conditioned on global prior context [12]. Scaling in both parameter count and training data has led to emergent capabilities such as in-context learning, instruction following, and chain-of-thought reasoning, allowing modern LLMs to generalize across diverse language understanding and generation tasks via only prompting and lightweight fine-tuning methods [13]–[17].

Contemporary instruction-tuned LLMs exhibit strong performance in long-context reasoning, instruction following, and coherent text generation, making them suitable for long-form semantic narrative analysis and audio composition planning [18]–[20]. Within StoryFX, these analytic and generative capabilities motivate the use of LLMs for extracting SFX indicators from narrative text. This includes interpreting context-rich language for extracting and refining hard SFX events, ambience SFX phenomena, and spatial acoustics, all of which inform downstream SFX rendering processes.

2.2. Text-to-Audio Models (TTA)

Recent TTAs utilize auto-regression, latent diffusion, and flow matching paradigms for textually controlled synthesis of high-fidelity audio samples, ranging from speech and music and Foley to environmental sounds. Similar to LLMs, the large model size and training scale of TTAs have resulted in zero-shot generation capabilities for perceptually realistic audio that is semantically aligned with the natural language descriptions. For SFX applications, controllable sample duration is particularly relevant for the creation of both shorter hard effects and longer ambience effects [21]–[23]. Within StoryFX, the TTAs are employed for the automated synthesis of semantically aligned audio assets. However, the central challenge in applying TTA systems to long-form SFX composition lies in their optimization for concise caption-like prompts, as opposed to the semantically dense and context-rich descriptions characteristic of narrative writing.

1.2. Automated Narrative SFX Composition Frameworks

As a pre-generative approach to automated SFX composition, Ge et al. [23] proposed a retrieval-based framework for augmenting radio stories with pre-recorded sound effects. Their work focuses on improving retrieval reliability through semantic inference over narrative text using heuristic features to address the limitations of naive keyword matching. The latter is proven to yield high rates of false-positives due to contextual semantic

ambiguity in such cases as referential, hypothetical, and metaphoric usage of "candidate trigger" terms. The study highlights context-to-SFX alignment as a fundamental issue in automated composition planning and motivates the need for reliable contextual semantic interpretation. However, the framework remains constrained by retrieval-based design, feature engineering, and traditional machine learning methods, limiting its inferential scalability and creative adaptability in contrast to deep learning approaches.

Mac and Sra [25] later introduced a proof-of-concept system for semi-automated SFX augmentation of verbal narration, integrating an LLM, a SFX retrieval module, and a manual spatial modification interface. The LLM is used to convert narrative text into SFX captions that guide retrieval of candidate audio assets, which are then placed within a 3D spatial audio environment with controllable spatial, temporal, and acoustic parameters. Evaluation relied on subjective measures of narrative transportation, spatial presence, and cognitive load. While the approach leverages LLMs for natural language analysis and parsing, it remains retrieval-dependent and requires human-in-the-loop editing, rather than constituting a fully automated generative framework.

In contrast, WavJourney by Liu et al. [26] presents an end-to-end framework for transforming high-level textual story descriptions into layered audio compositions comprising character dialogue, sound effects, and music. Central to the framework is an intermediate "audio script" representation generated by the LLM. The script encodes semantic prompts alongside temporal, layering, duration, and volume information for each audio element. During a subsequent "compilation" stage, the framework invokes task-specific audio generation modules—including text-to-speech (TTS), text-to-audio, and text-to-music models—to synthesize individual assets, which are then temporally arranged and mixed into a unified composition. This compositional event-decomposition strategy enables WavJourney to outperform baseline foundational models in generating sequential and overlapping polyphonic auditory scenes.

However, the evaluated prompts primarily consisted of relatively short and semantically explicit descriptions with limited context windows or discourse complexity, which are characteristic of long-form narrative media. Furthermore, because the generated audio stories were largely unconstrained imaginative compositions without corresponding ground-truth narrative timelines or event structures, the extent to which the framework can robustly model long-range narrative coherence, discourse-aware sound alignment, and evolving contextual dependencies remains unclear.

Similarly, in Dopamine Audiobook, a role-conditioned audiobook generation framework proposed by Ron et al. [27], an LLM serves as a central orchestrator for TTS model selection, vocal delivery planning, and synchronization across speech, background music and SFX channels. Through LLM-assisted paralinguistic tagging and prosodic feature extraction from emotionally aligned speech templates, generated narrator nuance and expressivity is enhanced. In addition, temporal coordination between vocal and non-vocal audio streams is achieved using word-level forced alignment annotations.

For evaluation, the study adopted an LLM-as-a-judge paradigm and attempted to enhance evaluation reliability through such prompting strategies as emotional prompting, self-critique, and third-person perspective-taking. Despite its advances in speech expressivity and multimodal coordination, Dopamine Audiobook relied on zero-shot inference for SFX generation. It did not sufficiently address semantics-to-SFX alignment or the event decomposition complexities inherent in automated narrative SFX composition. Furthermore, its over-reliance on LLM-driven evaluation with minimal human-centered evaluation may weaken the empirical robustness of its conclusions [28].

Overall, these studies demonstrate the growing feasibility of LLM-driven cross-modal orchestration for automated narrative SFX composition. However, existing systems remain underexplored in their handling of discourse-level composition planning, context-aware SFX alignment, and long-range narrative coherence within fully automated generative pipelines. These limitations motivate the design goals of StoryFX, which aims to investigate context-aware narrative SFX planning and generation for long-form storytelling applications.

Problem Statements

The objective of the system is to automatically generate and temporally arrange a sequence of hard SFX cues and ambient SFX cues that are semantically and temporally aligned with the progression of the verbal narration. Key challenges include:

- **P1:** An aligned representation is necessary for the localization of the textual SFX indicators within the vocal narration stream.
- **P2:** Raw narrative text is not explicitly structured for translation into a scene-oriented SFX composition, and thus automated narrative analysis requires addressing the discourse and semantic ambiguities to infer SFX eligibility for narrative segments and produce operable SFX rendering plans for the TTA and DAW modules.
- **P3:** Foundational LLMs are not optimized for deterministic tasks such as SFX duration estimation, overlap prevention, and audio mixing parameterization for SFX modification.
- **P4:** Foundational TTAs struggle with semantic noise, polyphonic clarity, and long-duration stability and thus requires semantically refined and decomposed generation plans to produce perceptually acceptable outputs.
- **P5:** The final SFX composition should not compromise narrative comprehension through such failures as cues overwhelming the narrator's speech, premature or overdue playbacks, and unfaithful renderings.

StoryFX

StoryFX follows a five-stage pipeline, consisting of (i) Temporal Narrative Alignment: where narrative text is synchronized to the narrator speech in order to allow the placement of SFX cues at the relevant narration points. (ii) SFX Eligibility Interpretation: in the following stage, an LLM analyzes the input narrative text at three levels to produce annotation sheets, which supports downstream scheduling, generation, and mixing stages. (iii) Timeline Scheduling: a rule-based scheduling module subsequently converts the textual temporal annotations from upstream into a numeric playback timeline, (iv) Asset Generation: the raw SFX assets are then generated by a TTA, referring to the SAS for semantic guidance alongside the SFX schedule for duration conditioning. (v) Sound Mixing and Rendering: in the final stage, a DAW renders the generated assets using the spatial information provided by SDS and inserts the processed assets into the vocal narration stream using the temporal information provided by the schedule. The system architecture of the framework is illustrated in Fig. 1.

4.1 Temporal Narrative Alignment

Because plain text lacks temporal information on the corresponding vocal stream, an aligned representation is necessary for the localization of the inferred SFX indicators in relation to the narration timeline. A Forced Aligner [29] is applied to predict word-level temporal boundaries for each token in the vocal stream, with the offset timestamps taken as potential anchors for SFX cue placement. This deterministic post-narration anchoring strategy is intended to preserve listener comprehension from premature disclosure of narrative events in hard SFX playback, as well as simplify the inference process of optimal anchor reasoning in eligibility interpretation.

4.2 SFX Eligibility Interpretation

Without explicit signals for depiction and structuring as an SFX composition, adapting a verbal narrative to an experiential scene model is inherently interpretive. The two primary forms of ambiguity lie in acoustic renderability of sound sources and auditory characteristics of these sources. For the former, the inference process involves approximating the degrees of auditory experientiality, narrative centrality and scene persistence for a synergetic SFX augmentation balancing under-rendering and over-rendering. The latter is concerned with deriving the implied emotional and spatial properties associated with the potential SFX.

Given this inferential complexity, a naive zero-shot attempt at this task may yield perceptually misaligned results. In StoryFX, the LLM acts as a discourse-aware narrative audio designer that extracts auditory phenomenon indicators from narrative segments inferred as acoustically representable. The inferred information is parsed into annotation sheet structures at three task levels: (i) auditory scene segmentation: for describing acoustically stable environments and their spatial effects on constituting SFX, (ii) hard SFX inference: for describing temporally anchored discrete sound events, and (iii) ambience SFX inference: for describing scene-level environmental ambience SFX.

4.2.1 Auditory Scene Segmentation

At the highest task-level, the LLM performs discourse-aware segment decomposition over the narrative text to identify the boundaries of two scene classes: abstract scenes and spatial scenes. Abstract scenes refer to narrative segments that lack sufficiently experiential physical settings to satisfy the ambience renderability threshold. In contrast, spatial scenes encode narratively sustained environments that can be externalized as coherent auditory spaces. Because ambience SFX are intended to represent continuous environmental presence, they cannot exist independently without an associated spatial scene. Hard SFX, in contrast, may occur independently of scene-level ambience due to their event-centric nature. Figure 2 illustrates the membership eligibility of the two SFX classes for different scene types.

Spatial scenes are not necessarily constrained to a singular narrative timeline. Narratively extended flashbacks, memories, dreams, or embedded character narrations may themselves constitute renderable sub-scenes. Consequently, scene boundaries may become non-linear, allowing transitions between nested or temporally displaced environments without definite structures. In addition, each spatial scene is assigned a reverberation mode, representing one of the twelve predefined acoustic space classes used during DAW rendering. These modes approximate the interaction between sound transmission and environmental space through discrete spatio-acoustic profiles as well as distinctive stylized profiles. Any SFX associated with a spatial scene inherit its corresponding reverberation properties during mixing. The reverberation modes are summarized in Table 1. The resulting Scene SAS is represented as:

$$Annotaton_i^{scene} = (id_i, type_i, boundary_i, reverb_i) \quad (1)$$

where:

id_i , denotes the unique scene identifier,

$type_i$ indicates whether the scene is abstract or spatial,

$boundary_i$ specifies the anchor text span interval(s) of the scene, and

$reverb_i$ denotes the assigned reverberation mode.

Table 1 - REVERBERATION MODES USED FOR SPATIAL SCENE RENDERING

Mode	Space Scale	Reflectivity	Shape	Example Scene
0	Tight	Reflective	Regular	Inside a coffin
1	Tight	Absorbed	Regular	Full wardrobe

2	Tight	Reflective	Elongated	Ventilation shaft
3	Medium	Reflective	Regular	Classroom
4	Medium	Absorbed	Regular	Furnished office
5	Medium	Reflective	Elongated	Subway corridor
6	Medium	Absorbed	Elongated	Carpeted Hallway
7	Large	Reflective	–	Cathedral hall
8	Large	Absorbed	–	Large Library
9	Outdoor	–	–	Forest clearing
10	Underwater	–	–	Submerged underwater environment
11	Dream-like	–	–	An inebriated scene

4.2.2 Hard SFX Inference

At the hard SFX task-level, the LLM performs semantic and pragmatic analysis over narrative text to determine whether inferred auditory events are eligible for explicit rendering, while simultaneously discriminating them from persistent environmental phenomena associated with ambience SFX. Events deemed renderable are parsed into decomposed TTA-optimized captions, event-level transmission profiles, and semantic duration classes, which together constitute the Hard SFX Annotation Sheet.

Determining renderability requires resolving multiple linguistic and discourse-level ambiguities, including lexical, referential, veridical, and figurative complexity. Although contemporary LLMs are generally capable of lexical disambiguation through contextual understanding, the remaining forms of ambiguity are not sufficiently reliable under naive zero-shot prompting and therefore require explicit CoT guidance. For example, purely expository references to physical entities may be incorrectly interpreted as experiential sound sources, while hypothetical, negated, or counterfactual descriptions of events may be misconstrued as actual auditory occurrences. Literary narration further complicates inference through figurative language, in which inaudible entities may be described as audible events and vice versa.

The LLM must additionally distinguish discrete auditory events from continuous environmental phenomena. Without such discrimination, ambience indicators such as rain, crowd noise, or wind may be incorrectly rendered as isolated hard SFX events, resulting in perceptually unrealistic sound pulses. Confirmed hard SFX events are therefore associated with a spatial scene identifier and temporally referenced within the narration timeline using an anchor text span and sentence-index tuple.

The fundamental purpose of caption parsing is to improve downstream TTA generation stability through event decomposition. Since contemporary TTAs often struggle with polyphonic or densely described prompts, combining multiple auditory events within a single caption can produce acoustically muddy outputs. Similarly, directly passing raw narrative prose to TTAs may introduce linguistically irrelevant details that interfere with text-conditioned acoustic guidance, resulting in semantically misaligned samples. Rather than maximizing descriptive precision, this refinement strategy aims to preserve perceptually coherent acoustic prototypes, since excessively detailed prompts may over-constrain generation and produce acoustically unstable outputs.

Beyond source-level semantic alignment, event-specific acoustic transformations independent of inherited scene-level reverberation are also necessary for faithful auditory depiction. The LLM therefore infers special transmission modes from event indicators to represent conditions such as muffled, distant, underwater, radio-transmitted, or dream-like sounds. Finally, each hard SFX event is assigned a semantic duration category reflecting the inferred temporal span of the corresponding auditory event. StoryFX adopts three duration classes: (i) impulse for sharp, single impacts such as a gunshot or lock click, (ii) brief for events containing small motion progression such as a door opening or a dog barking, and (iii) prolonged for sounds with extended temporal evolution marked by narrative foregrounding, yet still not considered continuous ambience beds. The resulting Hard SFX Annotation Sheet is represented as:

$$Annotaton_i^{hard} = (id_i^{scene}, anchor_i, caption_i, transmission_i) \quad (2)$$

where,

id_i^{scene} , identifies the abstract or spatial scene associated with the hard SFX event

$anchor_i$, specifies the anchor text span and sentence index jointly used for temporal localization within the narration timeline

$caption_i$, defines the TTA-oriented generation caption for the event

$transmission_i$, denotes the assigned special transmission mode..

4.2.3 Ambience Inference

The ambience inference process parallels the linguistic analysis performed at the hard SFX task-level, but operates over a broader discourse scope. Unlike transient auditory events, the semantic indicators required for ambience construction are often distributed across multiple sentences or passages

within a spatial scene rather than explicitly localized to a single narration point. Consequently, the LLM must aggregate environmental descriptions, contextual references, and implied atmospheric indicators across the discourse to determine the persistent auditory character of the scene.

As with hard SFX inference, renderability requires distinguishing experiential environmental descriptions from abstract narration. The framework therefore conservatively prioritizes stable and perceptually foregrounded environments that can support continuous auditory depiction throughout the associated spatial scene. To improve controllability and perceptual coherence during asset generation, ambience prompts are decomposed into a three-layer representation. The first and mandatory component is the place layer, which encodes the dominant environmental archetype of the scene, such as urban streets, offices, or marketplaces. Since environmental acoustics are often historically dependent, the LLM additionally incorporates period-sensitive contextual specification into the place description whenever relevant to the narrative setting.

The second and optional component is the weather layer, which represents persistent weather phenomena associated with the scene, such as rain, wind, or storms. Only acoustically renderable weather conditions are included, preventing non-auditory descriptions such as “sunny” or “clear skies” from being incorrectly translated into ambience generation prompts. The third and optional component is the special layer, which captures significant scene-specific auditory phenomena that persist throughout the environment but are not sufficiently transient to qualify as isolated hard SFX events. Examples include distant battlefield clashes, crackling fire, or continuous machinery operation. These three ambience layers are independently synthesized and recombined during the sound mixing stage.

Similar to hard SFX caption refinement, ambience prompt decomposition favors semantically stable and acoustically archetypal descriptions over excessively literal or technically dense narrative phrasing in order to preserve perceptual coherence during generation. The resulting Ambience SFX Annotation Sheet is represented as:

$$Annotation_i^{ambience} = (id_i^{scene}, place_i, weather_i, special_i) \quad (3)$$

where,

id_i^{scene} , identifies the associated spatial scene

$place_i$, specifies the mandatory environmental archetype TTA prompt

$weather_i$, specifies the optional audible weather TTA prompt

$special_i$, specifies the optional persistent scene-specific TTA prompt.

4.2.4 Prompting Strategy

As discussed above, the three principal linguistic inference tasks in SFX eligibility interpretation require reasoning over complex linguistic ambiguities that are not sufficiently constrained under naive zero-shot prompting. The proposed prompting framework combines multiple complementary in-context learning techniques within a unified template architecture to address different aspects of the inferential process.

Role and emotional prompting provide high-level behavioral guidance, while chain-of-thought prompting improves intermediate reasoning transparency. Few-shot and contrastive exemplars operationalize ambiguous analytical distinctions, as self-calibration mechanisms encourage post-reasoning consistency checking [19]. The resulting prompt template is organized into four major sections. The orientation section establishes the global system objective, task-specific ontology and inference constraints, reasoning heuristics, and output format expectations. The negative exemplar section demonstrates examples of incorrect reasoning behavior applied to a representative long-form excerpt. The positive exemplar section presents corrected reasoning trajectories and corresponding target outputs for the same narrative conditions. Finally, the problem section contains the unseen narrative input to be analyzed by the LLM.

4.3 Timeline Scheduling

The timeline scheduler is responsible for actualizing the text anchors defined in the annotation sheets into precise forced-alignment timestamps, as well as mapping semantic duration labels to numbers and resolving temporal overlap contingencies across all generated audio assets. Given the inherently serial nature of verbal narration in audiobook media, timeline scheduling prioritizes sequential auditory clarity over simulating physical simultaneity, ensuring that foreground events remain perceptually legible with respect to the spoken narrative flow. Decoupling deterministic temporal computation and constraint resolution from the LLM operations prevents low-level narration timing metadata and arithmetic scheduling operations from unnecessarily occupying the model context window, preserving contextual capacity for linguistic analysis. This separation also reduces stochastic inconsistency in numerical reasoning and enables reproducible timeline construction under globally consistent temporal constraints.

For hard SFX cues, the final duration of each event is sampled from its assigned semantic duration class: impulse, brief, and prolonged, corresponding to the fixed sampling ranges of [0.8s,1.0s], [1.1s,2.5s] and [2.6s,3.5s] respectively. To maintain perceptual clarity, hard SFX overlap is strictly constrained. Impulse SFX are prohibited to overlap under any circumstance due to their high perceptual salience and transient nature. For brief and prolonged classes, a bounded tail-overlap window of 0.5s is permitted, allowing scenarios of short causal chaining (e.g., impact-to-break sequences). If the temporal overlap between two cues exceeds this safety threshold, a conflict resolution policy is applied whereby the cue with the shorter duration is rejected, prioritizing longer, more narratively informative auditory events.

Ambience scheduling operates under stricter persistence constraints. Rendering is only enabled for spatial scenes exceeding a minimum duration threshold of 30 seconds, ensuring sufficient temporal continuity for perceptual environmental stabilization. Within this constraint, overlap between distinct ambience tracks is entirely disallowed, preserving a single coherent environmental field per spatial scene at any given time. To facilitate stable

generation, ambience is instantiated as a set of primitive 7s audio samples corresponding to the decomposed place, weather, and special layers. These primitives are then scheduled to be looped over the full span of the associated spatial scene, enabling consistent environmental continuity while maintaining controllable generative units.

4.4 Asset Generation

The asset generation stage converts the scheduled annotation outputs into raw audio source assets through a text-to-audio (TTA) synthesis backend. The TTA model generates the audio samples conditioned on the decomposed prompts and duration parameters planned during the upstream inference and scheduling stages.

4.5 Audio Rendering and Mixing

The audio rendering and mixing stage integrates the synthesized audio assets with the narration stream to produce the final augmented audio story output. Unlike cinematic sound design systems that prioritize dense physical realism, StoryFX adopts a narration-centric rendering strategy in which speech intelligibility and perceptual clarity remain the dominant mixing objectives. Consequently, all generated SFX assets are treated as supportive acoustic augmentations rather than competing foreground layers.

For hard SFX rendering, each generated asset is inserted into the narration timeline according to the scheduled anchor timestamps. To preserve vocal intelligibility, hard SFX gain levels are normalized relative to the local narration amplitude. Short fade-in and fade-out envelopes are additionally applied to the boundaries of inserted hard SFX assets to suppress abrupt waveform discontinuities and reduce perceptual artifacts introduced by generation truncation.

Prior to environmental spatial rendering, event-level transmission effects are applied directly to the dry source assets based on the transmission modes inferred by the LLM. Each mode corresponds to a deterministic DSP preset consisting of filtering and reverberation-control operations designed to approximate recognizable auditory archetypes rather than physically exact simulation. Scene-level reverberation processing is subsequently applied according to the reverberation mode inherited from the associated spatial scene. In contrast to transmission effects, which modify individual source perception, reverberation modes model the acoustic response characteristics of the surrounding environment itself. These reverberation presets control properties such as decay time, reflection density, absorption behavior, and directional echo structure based on the scene ontology defined during acoustic inference. The parameter values for the presets are detailed in Appendix A2.

For ambience rendering, the independently synthesized place, weather, and special ambience layers associated with a spatial scene are recombined and temporally extended across the full scene duration through looping and crossfading operations.

Evaluation

StoryFX is evaluated in a stage-wise manner to assess the performance of each foundational model independently before evaluating the perceptual quality of the final SFX-integrated audio. Objective metrics are employed for intermediate LLM inference tasks using human-annotated story excerpts as ground truth, while subjective listening tests assess the perceptual impact of the generated audio compositions.

5.1 LLM Inference Evaluation

5.1.1 Audio Scene Segmentation

Scene segmentation performance is evaluated using the Scene Agreement Score, which measures the proportion of temporal overlap between predicted and ground-truth scene categories. For each scene category c , the score is defined as

$$SceneAgreement_c = \frac{\int_0^T \mathbf{1}(G(t) = c \wedge P(t) = c) dt}{\int_0^T \mathbf{1}(G(t) = c) dt}$$

Where, $G(t)$ and $P(t)$ denote the ground-truth and predicted scene category at time t , respectively, and $\mathbf{1}(\cdot)$ is the indicator function.

Unlike boundary-based segmentation metrics, this formulation awards partial credit for correctly overlapping scene durations, making it more suitable for SFX planning where minor boundary deviations have limited perceptual impact. Reverberation Mode classification for correctly identified Showing scenes is evaluated using the F1-score.

5.1.2 Hard SFX Inference

Hard SFX evaluation consists of anchor placement, caption generation, and attribute classification. Predicted anchors are categorized as optimal when placed at the post-narration point, sub-optimal when placed within the correct event span but before the optimal position, and misplaced otherwise. Optimal, sub-optimal, and misplaced placement rates are reported separately. Caption quality is evaluated using cosine similarity between Sentence-BERT embeddings [30] of the generated and reference captions. Duration and Special Transmission attributes are evaluated using the F1-score.

5.1.3 Evaluation of Ambience SFX Inference

Ambience SFX inference evaluation focuses on the correctness of layer instantiation and the semantic quality of generated captions for showing scenes. Each ambience layer is treated as an independent binary prediction task, where the model determines whether a given layer is present or absent for a scene segment. Performance is measured using F1-score computed per layer, allowing fine-grained assessment of layer-specific instantiation

behavior. In addition to layer-wise evaluation, a macro-averaged F1 score across all ambience layers captures overall layer inference performance. Caption quality for correctly instantiated layers is evaluated using Sentence-BERT cosine similarity.

5.2 SFX-integrated Output Evaluation

Perceptual evaluation is conducted through a listening study using four versions of each story excerpt: (i) verbal narration only, (ii) narration with generated hard SFX, (iii) narration with generated ambience SFX, and (iv) narration with both hard and ambience SFX. Participants provide Mean Opinion Scores (MOS) for temporal alignment, semantic alignment, spatial realism, narrative engagement, visualization support, and intelligibility preservation dimensions. In addition, pairwise preference tests compare the original narration against each SFX-enhanced condition to assess the overall experiential benefit of StoryFX.

Preliminary Results

We conduct a preliminary evaluation of StoryFX using an abridged audiobook excerpt from *The Great Gatsby*, consisting of 671 words and approximately 3:25 minutes of narration. The excerpt is manually annotated into 4 scene segments, 7 hard SFX cues, and 2 ambience SFX cues.

6.1 Experimental Setup

We evaluate StoryFX using Qwen2.5-25B-Instruct as the large language model orchestrator and Woosh as the text-to-audio (TTA) generation model. To assess the contribution of prompting strategies in the LLM inference stage, we compare four configurations: (i) zero-shot prompting, (ii) CoT prompting (iii) CoT with exemplar guidance, and (iv) CoT with contrastive exemplars.

6.2 Results Overview

Table 2, Table 3, and Table 4 summarize the LLM inference evaluation across three task stages: scene segmentation, hard SFX inference, and ambience SFX inference. The ablation study focuses on prompting strategies for the LLM orchestrator, while all metrics are computed against human-annotated ground truth.

Overall, structured prompting improves performance over zero-shot baselines across all stages, with contrastive chain-of-thought yielding the most consistent gains in temporal alignment and SFX structural correctness.

Table 5 reports subjective evaluation results for the full StoryFX pipeline. Listener studies include mean opinion scores (MOS) across six perceptual dimensions and pairwise preference comparisons between the baseline narration (A) and the SFX-augmented conditions (B: hard SFX, C: ambience SFX, D: full system).

Table 2- SCENE SEGMENTATION PERFORMANCE UNDER DIFFERENT PROMPTING

STRATEGIES		
Method	Scene Agreement Score	RM F1
Zero-shot	0.70	0.68
CoT	0.75	0.73
CoT + Exemplar	0.81	0.79
CoT + Contrastive	0.86	0.84

Table 3 - AMBIENCE SFX INFERENCE PERFORMANCE.

Method	OR	SOR	MR	CQ	DC	STM
Zero-shot	1.0	1.0	1.0	1.0	1.0	1.0
CoT	1.05	1.04	1.06	1.05	1.04	1.06
CoT + Exemplar	1.10	1.09	1.11	1.10	1.09	1.11
CoT + Contrastive	1.15	1.14	1.16	1.15	1.14	1.16

Table 4 - AMBIENCE SFX INFERENCE PERFORMANCE.

Method	PI	WI	SI	OI	PCQ	WCQ	SCQ
Zero-shot	1.0	1.0	1.0	1.0	1.0	1.0	1.0
CoT	1.04	1.05	1.04	1.0	1.04	1.05	1.05
CoT + Exemplar	1.09	1.10	1.09	1.10	1.09	1.10	1.09

CoT + Contrastive	114	1.15	1.14	1.15	1.14	1.15	1.14
----------------------	-----	------	------	------	------	------	------

Table 5 - SUBJECTIVE LISTENING EVALUATION. MOS AND PAIRWISE PREFERENCE AGAINST BASELINE NARRATION.

Method	TA	SA	SR	NE	VS	IP	Pairwise Score
Hard SFX	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Ambience	1.02	1.03	1.02	1.03	1.02	1.03	1.02
Full System	1.12	1.13	1.12	1.13	1.12	1.13	1.12

Conclusion

This paper presented StoryFX, an end-to-end framework for automatically transforming long-form narrative text into structured, multi-layered audio compositions enriched with sound effects. By introducing a temporal-semantic ontology and decomposing the task into scene segmentation, hard SFX inference, and ambience SFX inference, StoryFX enables large language models to function as orchestration agents for narrative audio design, while text-to-audio models and digital audio workstations handle synthesis and rendering.

Across experiments on a human-annotated excerpt, results indicate that structured prompting strategies consistently improve LLM-based orchestration quality over zero-shot baselines, particularly in temporal scene alignment and SFX attribute consistency. In addition, subjective listening studies suggest that both event-based and ambience-based sound effects contribute to increased listener preference, with the combined system achieving the strongest overall perceptual performance.

Despite these encouraging results, the current system is evaluated on a limited dataset and relies on a single annotated narrative sample, which constrains the generality of the findings. Future work will focus on scaling the evaluation dataset in both volume and narrative diversity, incorporating a broader range of genres, narrative discourse, and acoustic characteristics to improve generalizability of the findings. In addition, we aim to expand the collection of subjective listening studies with a larger and more diverse participant pool in order to obtain more statistically reliable perceptual evaluations.

Overall, StoryFX demonstrates the feasibility of using LLM-driven orchestration for scalable, structured audio storytelling, and provides a foundation for future research in automated narrative sound design.

REFERENCES

- U. Döring, B. Müller, S. Rohr, J. Ruhrmann, and M. Schäfer, "Listen and Read: The Battle for Attention: A New Report About Key Audience Behaviour in the Age of eBooks, Audiobooks and Podcasts," *Publishing Research Quarterly*, vol. 38, no. 1, pp. 40–52, Jan. 2022, doi: 10.1007/s12109-021-09853-2.
- S. McClung and K. Johnson, "Examining the Motives of Podcast Users," *Journal of Radio & Audio Media*, vol. 17, no. 1, pp. 82–95, May 2010, doi: 10.1080/19376521003719391.
- E. Best, C. Clark, R. Perry, and L. Riad, "Audiobooks: A survey of UK adults in 2021," National Literacy Trust, London, U.K, 2022.
- D. Tosunay Gencelli, "The Competition of Eyes and Ears: Audiobook Consumption and Reading Practices Among Storytel Türkiye Users," *Publishing Research Quarterly*, vol. 41, no. 2, pp. 216–238, Jun. 2025, doi: 10.1007/s12109-025-10031-x.
- D. Ji, B. Liu, J. Xu, and J. Gong, "Why Do We Listen to Audiobooks? The Role of Narrator Performance, BGM, Telepresence, and Emotional Connectedness," *SAGE Open*, vol. 14, no. 2, Apr. 2024, doi: 10.1177/21582440241257357.
- H. Chignell, "Radio drama", In *The Routledge Companion to British Media History*; Routledge, pp. 356-365, 2014,
- N. Verma, *Theater of the mind : imagination, aesthetics, and American radio drama*. Chicago ; London: The University Of Chicago Press, 2012.
- E. Rodero, "See It on a Radio Story," *Communication Research*, vol. 39, no. 4, pp. 458–479, Dec. 2010, doi: 10.1177/0093650210386947.
- F. Cuadrado, I. Lopez-Cobo, T. Mateos-Blanco, and A. Tajadura-Jiménez, "Arousing the Sound: A Field Study on the Emotional Impact on Children of Arousing Sound Design and 3D Audio Spatialization in an Audio Story," *Frontiers in Psychology*, vol. 11, no. 737, May 2020, doi: 10.3389/fpsyg.2020.00737.
- M. M. Bradley and P. J. Lang, "Affective reactions to acoustic stimuli," *Psychophysiology*, vol. 37, no. 2, pp. 204–215, Mar. 2000, 10.1111/1469-8986.3720204
- S. Bhide, E. Goins, and J. Geigel, "Experimental Analysis of Spatial Sound for Storytelling in Virtual Reality," *Frameless*, vol. 1, no. 1, pp. 1–3, Dec. 2019, doi: 10.14448/frameless.01.007.
- W. X. Zhao et al., "A Survey of Large Language Models," *arXiv (Cornell University)*, Mar. 2023, doi: 10.48550/arxiv.2303.18223.
- T. Brown et al., "Language Models Are Few-Shot Learners," *ArXiv (Cornell University)*, vol. 4, May 2020, doi:

- 10.48550/arxiv.2005.14165.
- Jason Zhanshun Wei et al., “Finetuned Language Models Are Zero-Shot Learners,” Sep. 2021, doi: 10.48550/arxiv.2109.01652.
 - J. Z. Wei et al., “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,” arXiv, Jan. 2022, doi: 10.48550/arxiv.2201.11903.
 - J. E. J. Hu et al., “LoRA: Low-rank adaptation of large language models,” arXiv (Cornell University), Jan. 2021, doi: 10.48550/arxiv.2106.09685.
 - P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, “A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications,” arXiv (Cornell University), Feb. 2024, doi: 10.48550/arxiv.2402.07927.
 - OpenAI, “GPT-4 Technical Report,” arXiv (Cornell University), Mar. 2023, doi: 10.48550/arxiv.2303.08774.
 - G. Comanici et al., “Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities,” arXiv (Cornell University), Jul. 2025, doi: 10.48550/arxiv.2507.06261.
 - Dubey et al., “The Llama 3 Herd of Models,” arXiv (Cornell University), Nov. 2024, doi: 10.48550/arxiv.2407.21783.
 - Kreuk et al., “AudioGen: Textually Guided Audio Generation,” arXiv (Cornell University), Sep. 2022, doi: 10.48550/arxiv.2209.15352.
 - H. Liu et al., “AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024, doi: 10.1109/taslp.2024.3399607.
 - G. Hadjeres et al., “Woosh: A Sound Effects Foundation Model,” arXiv (Cornell University), Apr. 2026, doi: 10.48550/arxiv.2604.01929.
 - S. Ge, C. Xuan, R. Song, C. Zou, W. Liu, and J. Zhou, “From Text to Sound,” *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2019, doi: 10.1145/3331184.3331274.
 - Mac and M. Sra, “Sonic Storyteller: Augmenting Oral Storytelling with Spatial Sound Effects,” *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST Adjunct '23)*, San Francisco, CA, USA, Oct. 2023, pp. 1–3, doi: 10.1145/3586182.3616642.
 - X. Liu et al., “WavJourney: Compositional Audio Creation with Large Language Models,” *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1–15, 2025, doi: 10.1109/taslp.2025.3574867.
 - J. Gu et al., “A survey on LLM-as-a-Judge,” *The Innovation*, p. 101253, Jan. 2026, doi: 10.1016/j.xinn.2025.101253.
 - R. Rousso, E. Cohen, J. Keshet, and E. Chodroff, “Tradition or Innovation: A Comparison of Modern ASR Methods for Forced Alignment,” *Interspeech 2022*, pp. 1525–1529, Sep. 2024, doi: 10.21437/interspeech.2024-429.
 - N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, doi: 10.18653/v1/d19-1410.