AudioXtend: Assisted Reality Visual Accompaniments for Audiobook Storytelling During Everyday Routine Tasks

Felicia Fang-Yi Tan
felicia.tanfy@gmail.com
Synteraction Lab
National University of Singapore
Singapore, Singapore

Peisen Xu
peisen@comp.nus.edu.sg
Synteraction Lab, IPAL
National University of Singapore
Singapore, Singapore

Ashwin Ram
ashwinram10@gmail.com
Synteraction Lab
National University of Singapore
Singapore, Singapore

Wei Zhen Suen
wei.zhen.suen@gmail.com
Synteraction Lab
National University of Singapore
Singapore, Singapore

Shengdong Zhao*
shengdong.zhao@cityu.edu.hk
City University of Hong Kong
Hong Kong, China
National University of Singapore
Singapore, Singapore

Yun Huang
yunhuang@illinois.edu
School of Information Sciences
University of Illinois at
Urbana-Champaign
Champaign, Illionois, United States

Christophe Hurter
christophe.hurter@enac.fr
ENAC, Université de Toulouse
Toulouse, France

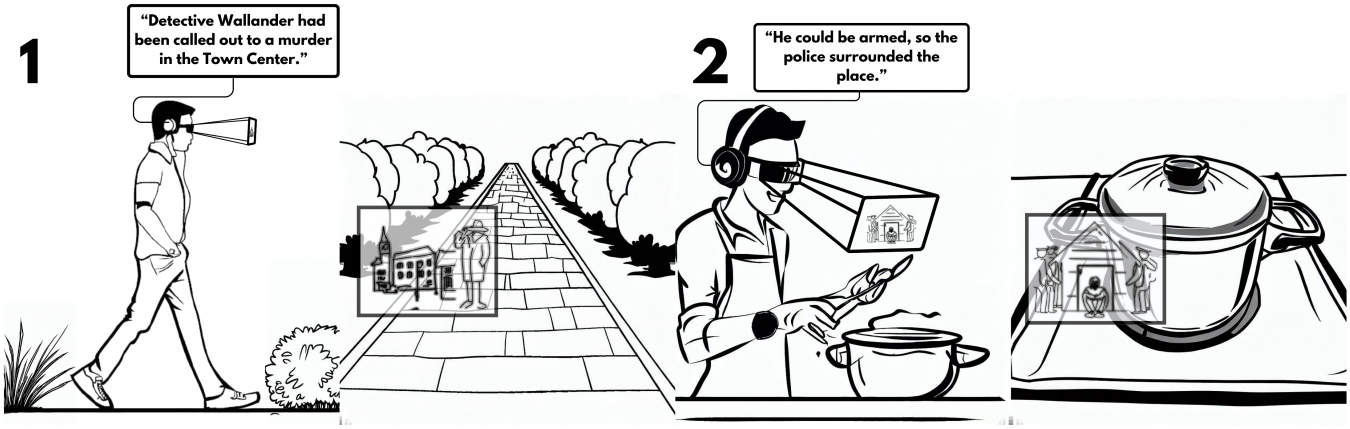

Figure 1: AudioXtend in action: Integrating glanceable visuals on optical see-through head-mounted displays (OHMDs) for enhanced incidental learning in multitasking contexts. (1) A third-person view and first-person perspective of a real-world walking activity where AudioXtend's AI-generated visuals and auditory narration synchronize for an improved learning experience. (2) Another third-person view and first-person perspective of a user cooking without disruption while using AudioXtend.

*Corresponding Author

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642514

ABSTRACT

The rise of multitasking in contemporary lifestyles has positioned audio-first content as an essential medium for information consumption. We present AudioXtend, an approach to augment audiobook experiences during daily tasks by integrating glanceable, AI-generated visuals through optical see-through head-mounted displays (OHMDs). Our initial study showed that these visual augmentations not only preserved users' primary task efficiency but also dramatically enhanced immediate auditory content recall by 33.3% and 7-day recall by 32.7%, alongside a marked improvement in narrative engagement. Through participatory design workshops

involving digital arts designers, we crafted a set of design principles for visual augmentations that are attuned to the requirements of multitaskers. Finally, a 3-day take-home field study further revealed new insights for everyday use, underscoring the potential of assisted reality (aR) to enhance heads-up listening and incidental learning experiences.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in ubiquitous and mobile computing**.

KEYWORDS

Smart-glasses, Optical See-Through Head-Mounted Displays, Assisted Reality, Incidental learning, Recall Enhancement, Visual Storytelling, Audiobook Augmentation, Heads-Up Computing

ACM Reference Format:

Felicia Fang-Yi Tan, Peisen Xu, Ashwin Ram, Wei Zhen Suen, Shengdong Zhao, Yun Huang, and Christophe Hurter. 2024. AudioXtend: Assisted Reality Visual Accompaniments for Audiobook Storytelling During Everyday Routine Tasks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/3613904.3642514

1 INTRODUCTION

Audiobooks have carved out a unique space within the realm of storytelling, offering a modern, digitized form of narrative experience. The art of narration enriches content with layers of tone, pacing, and emotion, offering an immersive experience that is distinct from text-based reading [47, 80]. Audiobooks have garnered immense popularity due to its ease of access and the convenience of hands-free and eyes-free consumption, often in multitasking scenarios [12, 70]. As many as 70% of audiobook listeners consume audio-first information while multitasking, such as while doing chores or exercising [93]. This underscores how the auditory channel has become a compelling means of content absorption in today's fast paced world.

Despite notable advantages, audiobooks are not without limitations in the learning context. Previous works have shown significantly lower levels of memory retention and comprehension when learning from audiobooks compared to traditional text [24]. Here, the concept of informal and incidental learning, as defined by Livingstone [64] and Schugurensky [101], is central to our approach. In contrast to formal learning which is typically structured and classroom-based, incidental learning takes place unintentionally, as a byproduct of some other activity [67]. For example, someone might learn historical facts while listening to a historical novel about a soldier on her car radio. Similarly, audiobooks are often consumed during everyday routine activities as they allow multitasking [46], providing a unique opportunity for incidental learning. Given this context of incidental learning during activities like commuting or cleaning, we aim for AudioXtend to leverage story-driven narratives (explained further in Section 2.2) to enhance the informal and incidental learning experience.

As suggested by Lee et al. [59], one contributing factor for poorer retention is the nature of audio as an invisible, intangible, and ephemeral medium that 'dissipates' as soon as narrated, providing no visual anchor for the listener. Prompted by these challenges, we asked: Is it possible to retain the multitasking advantages of audiobooks while enriching the learning experience? This motivated us to explore the idea of subtle and glanceable visual augmentations on Optical See-Through Head-Mounted Displays (OST-HMDs, OHMDs) as a supportive and complementary channel to audio-first content.

Past works have established the positive effects of illustrative visuals in facilitating learning, enhancing attention, memory and comprehension [29, 99], offering cognitive anchors that assist in information retention [86]. However, their usage alongside audiobooks can limit capabilities in the everyday multitasking context, presenting a dilemma: How can we combine visual augmentations with its auditory-base without compromising the primary task's performance?

To investigate this question, we introduce AudioXtend, a technique designed to augment auditory content with subtle, glanceable AI-generated visuals via OHMDs. By leveraging story-driven narratives from audiobooks as the primary modality channel, AudioXtend aims to enrich the user experience and enhance content absorption by adding synchronized illustrative visuals as a complement. It is worth noting that this proposal is conceptually distinct from video-based learning. Videos tend not to be designed for multitasking situations, often requiring full attention to continuous and persistent audio and visual components [83]. Audiobooks, however, are primarily consumed in situations where the auditory channel is more freely available, such as while exercising, commuting or performing household chores. The objective here is not to transform audiobooks into videos, but to augment them with selective visual illustrations that can be glanced at briefly to enrich engagement and retention without disrupting the multitasking context. This approach seeks to create an improved audiobook experience that maximizes the outcomes of incidental learning, without hampering the performance of primary tasks or diluting the core audiobook experience.

With this goal, we began in Study 1 by empirically evaluating a "naive" case of visual augmentations on memory recall and engagement, comparing it with audio-only content. In so doing, we established the potential contribution of such an audio-primary and visual-secondary design. Study 1 also revealed a variety of emerging factors which called for a more nuanced exploration of design factors for AudioXtend. Thus, we further extended our inquiry of the "what, when and how" by conducting participatory design workshops setup with everyday multitasking stations to yield crucial insights on the design space. Lastly, a take-home field study served to validate our framework's applicability and benefits in real-world settings. Our findings reveal that AudioXtend significantly enhances recall (immediate and 7-day) and engagement, while preserving the efficiency of users' primary tasks. These insights lay the foundation for reimagining audio-first experiences via assisted reality technology, suggesting that strategic visual augmentations could improve how we engage with auditory media in multitasking scenarios.

The contributions of this work include:

(1) An empirically-validated approach, AudioXtend, that leverages OHMDs to provide glanceable visual augmentations synchronized with audiobook content.

(2) A design space of effective visual augmentations for AudioXtend elicited through a hands-on participatory design approach involving individuals with expertise in digital art editing and AI-image generation.
(3) Design recommendations validated through a real-world field study, demonstrating the practical applicability and benefits of AudioXtend and its accompanying design guidelines in real-world incidental learning and leisure contexts.

2 RELATED WORK

This section outlines previous works that inform our approach, broken down into four key areas.

2.1 Audio-based Learning and Cognitive Theories

Audiobooks have gained prominence not just as a convenient method for consuming literature but also as an educational platform offering literacy improvement and vocabulary development [58]. Studies have shown that the combination of reading while listening can improve both reading fluency and comprehension, especially for younger or struggling readers [34, 105].

The dominant focus of HCI research on audiobooks, however, has been on enhancing user interface designs [59] and creating mood-augmenting environments [85, 112]. There remains a gap in explorations of the audiobook as potentially more than a unimodal medium. Specifically, little work has been done to explore how audiobooks could integrate additional channels of information, such as visuals, to create a richer learning experience.

Established cognitive theories in text-based learning offer valuable insights that are relevant to this extension. For example, Schallert [99] identified that illustrations help readers focus their attention on information in text and reorganize information into useful mental models. Paivio's "Conceptual Peg" hypothesis proposes that mental imagery, represented through illustrations or visuals, serves as "pegs" onto which verbal information can be "hooked" for easier storage and retrieval [86]. Paivio also highlights the integrative function of verbal and non-verbal systems where language can evoke mental imagery and vice-versa. Similarly, Mayer's Cognitive Theory of Multimedia Learning [71] shows that using both auditory and visual channels can reduce cognitive load, which is especially important in the multitasking context.

Supporting this, empirical tests like those by Jiang et al. underline the efficacy of multimedia glosses in vocabulary learning [53], demonstrating marked improvements in vocabulary recognition and retention. Orrantia's work further highlights the benefit of visual aids, especially for learners who might find it challenging to form mental representations from text alone [84]. Despite these promising results, no studies have been undertaken to apply them to the audiobook context. This requires moving beyond the perception that audiobooks are a strictly unimodal medium, and exploring multimodal design possibilities.

2.2 Story-driven Narratives in Audiobook Format

The choice to leverage story-driven narratives in AudioXtend is inspired by existing research on narrative and expository content. Presented and organized differently, many theorized differences in each category's potential for readers to retain and comprehend the information presented [39, 41]. A meta-analysis conducted by Wolfe and Mienko [119] investigating memory and comprehension of narrative and expository texts found that stories were more easily understood and better recalled than essays. Past research also suggests that narrative text better complements the audiobook format than expository text [105]. Expository texts, conventionally viewed as "informational" due to their factual and data-rich content, align better with the print medium. Print allows readers to control their pace, revisit complex sections, and engage in thorough analysis. This focused concentration and detailed processing is more effectively achieved through reading than listening. Conversely, narrative texts, characterized by their story-driven structure and focus on characters and plot, align well with the audiobook format. The sequential and time-bound nature of narratives complements the listening experience, allowing for a more immersive and emotionally engaging interaction with the content [94]. Audiobooks enhance the story's impact by bringing characters to life through voice and tone, thereby making the narrative more memorable and relatable [14, 87, 91]. This natural fit of narrative texts in audiobooks underpins our study's focus on visually augmented audiobooks. By leveraging the inherent strengths of audiobooks in presenting narratives, we aim to enhance the overall learning experience in the context of everyday activities where audiobooks are increasingly becoming a popular choice.

2.3 Multitasking in Everyday Settings

Research has repurposed the Activities of Daily Living (ADL) framework, which was originally conceived for patient rehabilitation, to categorize common daily tasks such as cooking and dressing [27]. This taxonomy is particularly relevant when understanding the interaction between incidental learning through audiobooks and everyday task engagement. Prior work has proposed a resource interaction model [125] that further elucidates this by considering how individuals allocate cognitive and physical resources across tasks. The model provides an inspiration to evaluate the suitability of adding visual augmentations to audiobooks depending on the task at hand.

This line of research aligns with developments in OHMDs and assisted reality technology, which shows promise for optimizing information processing in multitasking environments, given its see-through capabilities [38, 95]. They also enable hands-free operation, thereby reducing the fatigue associated with holding a device for extended periods [88]. These features position OHMDs as a compelling platform for on-the-go incidental learning, but also highlights the need to design content for effective multitasking without cognitive overload.

2.4 Glanceable Design in Assisted Reality for Learning

The concept of "glanceability," or the ease with which a user can quickly obtain the necessary information with a quick glance, has been explored in various contexts, from augmented reality (AR) to smartwatches and peripheral displays [11, 16, 48, 66, 68, 69, 79, 109].

In the realm of AR, Lu et al. examined glanceable information access methods on head-worn devices, emphasizing the importance of the user's ability to quickly obtain relevant data without disrupting ongoing tasks [66]. Similarly, Blascheck et al. evaluated glanceable visualizations on smartwatches, demonstrating the efficacy of quick, low-effort data comparisons [11].

In the context of multitasking and ADL, glanceability can significantly influence the success of information absorption. Peripheral displays like Sideshow and the Information Percolator have been designed to offer glanceable, ambient awareness of critical information [16, 48]. Matthews et al. extended this notion into the design principles for glanceable peripheral displays, providing a foundation upon which OHMD-based solutions like the proposed AudioXtend can build [68, 69].

Similar to research seen in AR and peripheral display research, the challenge we address is integrating glanceability into the AudioXtend system, to support incidental learning and effective multitasking without overwhelming the user's cognitive resources.

3 STUDY 1: COMPARING AUDIOBOOK WITH AUDIOXTEND

To examine the foundational case of whether visual augmentations serve to enhance the experience of consuming audiobooks, we employed illustrations that are based on preliminary parameters derived from early exploratory tests. This study seeks to answer the fundamental question: Are illustrations a useful supplement to audiobooks? On these grounds, we hypothesize the following:

- **H1** The audio+illustrations modality leads to better content recall immediately and after 7 days when compared to the audio-only modality. Incorporating visual cues should theoretically enhance the cognitive process by allowing for dual encoding (i.e. visual and auditory) [86], thereby improving recall.
- **H2** The audio+illustrations modality is rated as more engaging compared to the audio-only modality. Visual storytelling elements have been found in previous works to augment the emotional engagement and attentional focus of the user by enriching the narrative [42], thereby making the overall experience more immersive.
- **H3** There is no difference in primary task performance, perceived disruption to primary activity and perceived task load between the audio-only and audio+illustrations modality. Since the illustrations are designed to be subtle and glanceable, they are intended to complement rather than interfere with the primary task, thereby not increasing the overall cognitive load [117].

3.1 Participants

We recruited 12 participants (6 Females, 6 Males) from the university community. Their mean age was 23.8 (SD = 1.94) years. All were fluent in English, and spoke it either as their first or second language. All had prior experience listening to audio-first content (e.g. podcasts, audiobooks), but had never listened to the audiobook materials used in the study (refer to Materials section). Each participant was compensated ≈ \$7.5/h for their time. All studies were approved by the IRB of our institution, and an informed consent was obtained from every participant.

3.2 Measures

Our research, while touching upon multitasking effectiveness, primarily aims to address existing pain points of audiobook listeners. A key challenge for audiobook listeners, as highlighted in the introduction, is the reduced retention of content compared to reading printed books, especially since listening often occurs in multitasking contexts. We therefore measure that visual augmentations improve content recall without compromising multitasking performance. Narratives in audiobooks serve broader purposes beyond factual recall, and measuring the retention of story details serves as a proxy for other dimensions of engagement. This approach aligns with existing literature in psychology and learning science, where recall has been used to evaluate engagement with narrative content [2, 7, 123]. Our goal is also to improve subjective enjoyment and emotional impact of the stories. Thus, we included a subjective scale to evaluate overall engagement with the content.

Recall (Immediate and 7-day). We assessed both immediate and 7-day retention via a free recall test which required participants to "Write down all that you remember and understood of the content (be as descriptive as possible)", similar to that used by Ram et al. [89]. This gauges users' ability to retrieve information from memory without significant cues. Two independent raters marked the tests using a predefined marking scheme which awarded 1 point for correct content (names, objects, actions, events, concepts), descriptiveness (details on content), and comprehension (interpretation of context, emotions or motivations). Inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC). With an average measures ICC of 0.908 (95% CI [0.80, 0.97]), the reliability among raters was deemed high. Consequently, the final recall score for each participant was determined by averaging the scores from the two raters. Seven days post-experiment, participants received a prompt for the follow-up recall test in the same format.

Primary Task Performance. Performance was measured as a percentage of the participant's preferred speed in two tasks: walking and folding laundry. This metric indicates the relative slowdown a participant experiences when multitasking with the audiobook. To obtain the preferred speed, we first recorded the distance walked or number of completed laundry items folded in a 2 min "calibration" phase. The percentage of preferred speed (0-100%) was then obtained by dividing the speed of the primary task by the preferred speed.

Narrative Engagement Scale. Participants rated their engagement using a 7-point Likert scale that encompasses 4 constructs with 3 questions each: Narrative understanding, Attentional focus, Presence, and Emotional engagement. This is a well-established scale developed by [15], used in many past works such as [44, 62, 77, 104], and can be combined into a measure of overall engagement or divided into subscales that distinguish among different aspects of engagement.

Perceived Multitasking Experience. Participants rated the extent to which the audiobook disrupted their primary task using a single

7-point Likert scale question: How smooth was your experience of executing the primary task (walking navigation or folding clothes) while simultaneously engaging with the audiobook content or viewing illustrations?

NASA-TLX. The NASA-TLX [45] assessed participants' perceived task load, capturing the physical and mental effort exerted, through a rating system.

3.3 Materials

3.3.1 Audiobook Selection. We selected audiobooks with a narrative/story structure to ensure that illustrations could effectively complement and enhance the auditory experience. For greater ecological validity, tracks were sourced from real-life, well-produced audiobooks released by established publishing houses.

To mitigate potential confounding factors related to familiarity, we ensured, via a pre-experiment survey, that participants had neither read nor listened to the selected books. For the study, we extracted 5-minute excerpts from each:

(1) "Blood Work" by Michael Connelly (Read by Dick Hill, Audible)
(2) "Dogs of Riga" by Henning Mankell (Read by Dick Hill, Audible)
(3) "The Martian" by Andy Weir (Read by RC Bray, Podium Publishing)
(4) "The Ocean at the End of the Lane" by Neil Gaiman (Read by Neil Gaiman, Headline Books)

To further control for potential confounds related to comprehension difficulty, we selected audiobook tracks that exhibited similar levels of readability. All chosen tracks scored either 82 or 83 on the Flesch Reading Ease Test. This score indicates that the content is likely to be understood by someone with at least a 6th-grade education (approximately age 11).

3.3.2 AI-assisted Image Generation. For each 5-minute audiobook track, 14-16 images were generated to synchronize with the content. Based on feedback from early testing, this was the ideal range. With an iterative approach and the assistance of GPT-4 [1], a Large Language Model (LLM) from OpenAI, one experimenter identified sections within the 5-minute passage that held high relevance to content, emotional impact, significance within the overall passage, and potential for translation into visual representations.

Two other independent experimenters subsequently refined this initial draft. In cases of dispute, discussions were held to reach a consensus. Based on this refined draft, illustrations were generated using the AI Image generation tool, Microsoft Bing [72].

Each prompt provided to the AI began with "Simple outline drawing, no fills, with a black background…" We opted for a black background because, when displayed on OHMD hardware such as the Xreal Light (see Section 3.3.3), black color on screen appears transparent. This allowed only white outlines of the illustrations to appear superimposed on the external environment, creating a less obtrusive visual experience for the user.

The most apt and accurate illustrations were collaboratively chosen and synchronized with the audiobook track using a video editor. A second experimenter further refined this synchronization by testing the visuals on the OHMD. Based on early feedback, the placement of these visuals was deliberately off-center (to the left) and sized at 400 by 400 pixels (refer to first-person view in Figure 3). This design choice was made as it least disrupted the user's line of vision, ensuring the primary task remained unhindered while still being easily glanceable [21]. In addition, illustrations that included characters were labeled with the respective character names. This decision was also informed by early testing feedback, where participants indicated difficulty in discerning the spelling of certain character names based solely on auditory cues. The text labels were of 28 pt Inter font type.

On average, it took approximately 2 hours in total to create a full illustration set for each 5-minute audiobook track, though we acknowledge the potential for future work to automate this process further, thereby reducing the manual effort required (further discussed in Section 6.5). Figure 2 summarizes the AI-assisted image generation workflow.

3.3.3 Apparatus. Visuals were displayed on an Xreal Light OHMD [122], which places the display in the center of the user's line of sight. It has a 52-degree field of view and a stereoscopic display with a resolution of 1920 x 1080 pixels. In air casting mode, its screen is 115 inches at 3 meters. The visuals were pre-loaded onto an android phone which was mirrored to the OHMD. Audio was played through wireless headphones (Model: Bose QuietComfort 35 II) along with the OHMD apparatus.

3.3.4 Method. The study employed a fully-counterbalanced within-subjects design, structured as a 2x2 factorial, i.e. Modality x Primary Activity. We also counterbalanced the 4 audiobook tracks (Section 3.3.1) among conditions to ensure that any differences in audiobook content and order effects are mitigated.

Modality. This factor investigated the difference between co-presentation of visuals with audio versus an audio-only base. The two levels under this factor were thus Audio-only and Audio+Illustrations.

Primary Activity. This factor aligns with our focus on informal and incidental learning, where we aim to utilize idle moments of the day more effectively, not necessarily to create focused learning environments. It was introduced to generalize the findings to two common multitasking activities that mirror situations where audiobook consumption is both common and practical, allowing listeners to absorb content without dedicated attention. Our selection of primary activities was guided by the framework proposed by Salvucci et al. [96], which differentiates between "Highly Concurrent Task Performance" and "Sequential Interleaving of Tasks."

Indoor walking navigation: The concept of "Highly Concurrent Task Performance" aligns with tasks that demand cognitive resources to be constantly allocated to them, and interruptions or momentary lapses in attention can have immediate and noticeable consequences. Within this category, we chose indoor walking navigation as a primary activity. This task is not only commonplace but also necessitates continuous attention and physical movement.

The walking path chosen for this study is reminiscent of the 8-figure path, as utilized in previous research [50, 88], albeit at a longer distance of 123.4 meters for one round. Such a path is representative of real-world scenarios that demand more intricate motor skills than a simple straight path.

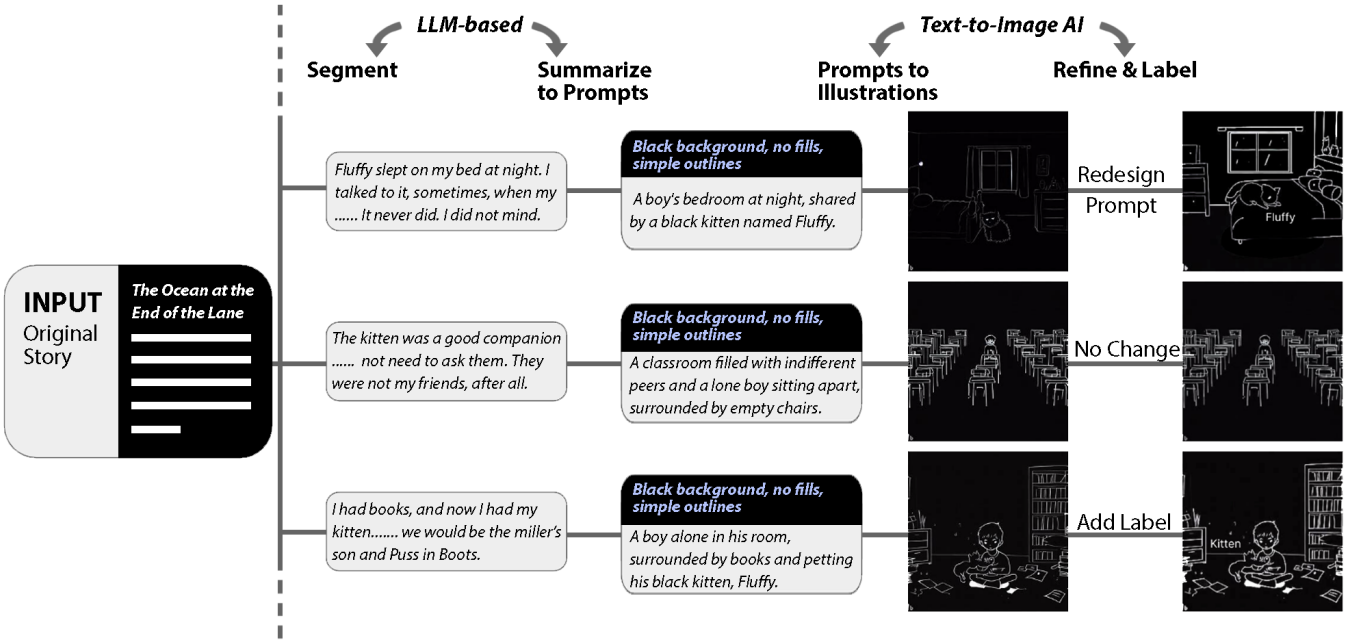

Figure 2: AI-assisted image generation workflow.

Folding laundry while standing: The "Sequential Interleaving of Tasks" as described by Salvucci et al., refers to tasks that are performed sequentially, often allowing pauses or diversions in attention that do not result in immediate negative consequences. Employed in prior research involving multitasking [51, 98], folding laundry is the representative task we chose as it allows for such flexible shifts in attention. It is also a common physical task laid out in the Activities of Daily Living taxonomy (refer to section 2.2).

In the Folding Laundry task, participants were positioned standing and were presented with laundry on a table, encompassing clothes like t-shirts, blouses, dresses and pants. They were instructed to fold these clothes in a sequential and consistent manner, following the Marie Kondo method [82], using both hands. To ensure variability, the sequence of these clothes was randomized after each trial. While the other primary activity of walking required participants to be mobile, this activity allows us to understand the effects of standing posture with varied hand movements.

3.3.5 Procedure. Upon arrival, participants signed a consent form and were introduced to a sample audiobook and the two primary tasks: walking on a designated indoor path and folding laundry. Before the main experiment, a 2-minute "calibration" phase was conducted where participants performed their primary task without any audiobook, so that we could obtain their preferred primary task speed (see Section 3.2). The main experiment involved listening to a 5-minute audiobook segment while performing a primary task, followed by a recall quiz and a post-experiment questionnaire. This process was repeated for three more audiobooks, with conditions counterbalanced. After all segments, a brief interview was conducted to gather qualitative feedback. One full experiment lasted approximately 1.5 hours. Seven days post-experiment, participants completed a follow-up recall quiz to assess long-term retention.

3.4 Results

A 2x2 within-subjects factorial ANOVA was used to analyze all data. Both Sphericity (Mauchly's test of sphericity, $p > 0.05$) and Normality (Shapiro-Wilk test, $p > 0.05$) assumptions were met for all.

3.4.1 Audiobook Recall.

Immediate. There was a significant main effect of Modality, $F(1, 11) = 14.02$, $p < 0.01$, $\eta_p^2 = 0.56$, indicating that the type of modality significantly affected recall performance, irrespective of the activity being performed.

The mean recall score was 33.3% higher in the Audio+illustrations condition ($M = 21.6$, $SD = 6.49$) compared to the Audio-only condition ($M = 16.2$, $SD = 7.91$).

After 7-days. There was also a significant main effect of Modality, $F(1, 11) = 5.24$, $p < 0.05$, $\eta_p^2 = 0.32$, indicating that the type of modality significantly affected recall performance, irrespective of the activity being performed.

The mean recall score was 32.7% higher in the Audio+illustrations condition ($M = 14.6$, $SD = 6.62$) compared to the Audio-only condition ($M = 11.0$, $SD = 7.35$).

While there was a drop in scores for both modality conditions after 7 days, as expected, the advantage of using Audio+Illustrations over Audio-only remained relatively consistent, offering statistically significant advantage, and decreasing by just 0.6% from the immediate test to the 7-day test.

On the other hand, we found no significant main effect of Activity or Activity X Modality interaction for both immediate and 7-day recall. This suggests that the influence of modality on recall performance did not differ between the two activities.

(a) (b) (c) (d)

Figure 3: Main activities performed in Study 1. (a) Participant walking and (b) respective first-person view. (c) Participant folding laundry and (d) respective first-person view.

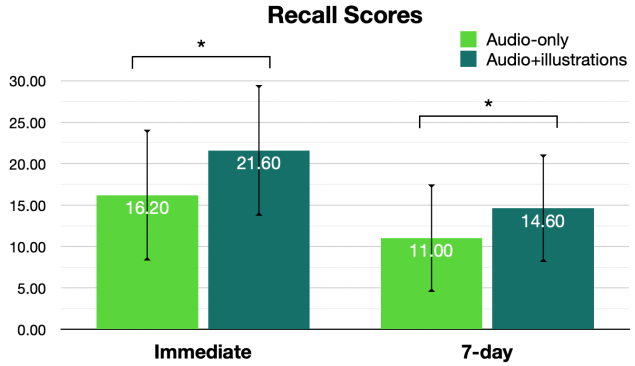

Figure 4: Mean scores for Immediate and 7-day recall, in the audio-only compared to the audio+illustrations modalities

3.4.2 Primary Task Performance. We found no significant main effect of Modality, Activity and Activity X Modality interaction, suggesting that none of these factors or interaction had a significant effect on the speed of task completion.

3.4.3 Narrative Engagement.

Narrative Understanding. There was a significant main effect of Modality, $F(1, 11) = 5.670$, $p < 0.05$, $\eta_p^2 = 0.340$, indicating that the type of modality significantly affected the narrative understanding, irrespective of the activity being performed. The mean rating for narrative understanding was 38.7% higher in the Audio+illustrations condition ($M = 4.28$, $SD = 1.03$) compared to the Audio-only condition ($M = 3.08$, $SD = 1.22$).

Attention Focus. There was a significant main effect of Modality, $F(1, 11) = 8.939$, $p < 0.05$, $\eta_p^2 = 0.448$, indicating that the type of modality significantly affected the attention focus, irrespective of the activity being performed. The mean rating for attention focus was 31.9% higher in the Audio+illustrations condition ($M = 4.05$, $SD = 1.09$) compared to the Audio-only condition ($M = 3.07$, $SD = 0.81$).

Narrative Presence. There was a significant main effect of Modality, $F(1, 11) = 5.837$, $p < 0.05$, $\eta_p^2 = 0.347$, indicating that the type of modality significantly affected the narrative presence, irrespective of the activity being performed. The mean rating for narrative presence was 19.6% higher in the Audio+illustrations condition ($M = 4.40$, $SD = 0.90$) compared to the Audio-only condition ($M = 3.68$, $SD = 1.00$).

Emotional Engagement. We found no significant main effect of Modality (Audio+Illustration: $M = 3.94$, $SD = 1.17$, Audio only: $M = 3.24$, $SD = 1.19$), Activity and Activity X Modality interaction, suggesting that none of these factors or interaction had a significant effect on emotional engagement.

Overall Engagement. There was a significant main effect of Modality, $F(1, 11) = 9.630$, $p < 0.05$, $\eta_p^2 = 0.467$, indicating that the type of modality significantly affected the overall engagement, irrespective of the activity being performed. The mean rating for the overall engagement was 27.6% higher in the Audio+illustrations condition ($M = 4.17$, $SD = 0.67$) compared to the Audio-only condition ($M = 3.27$, $SD = 0.76$).

For all dimensions of engagement above, we found no significant main effect of Activity, indicating that whether the participants were walking or folding laundry did not significantly affect their overall engagement. There was also no significant effect of the Activity X Modality interaction suggesting that the influence of modality on the overall engagement did not differ between the two activities.

3.4.4 Multitasking Experience. For the perceived main task effectiveness, there was no significant main effect for modality. However, a significant main effect of activity emerged, $F(1, 11) = 43.703$, $p < 0.001$, $\eta_p^2 = 0.799$. Specifically, participants rated walking ($M = 5.88, SD = 0.98$) as a 42.4% smoother task to execute than folding laundry ($M = 4.13$, $SD = 1.51$).

There was no significant interaction between Activity and Modality.

3.4.5 NASA-TLX Cognitive Load. We found no significant main effect of Modality, Activity and Activity X Modality interaction, suggesting that none of these factors or interaction had a significant effect on cognitive load.

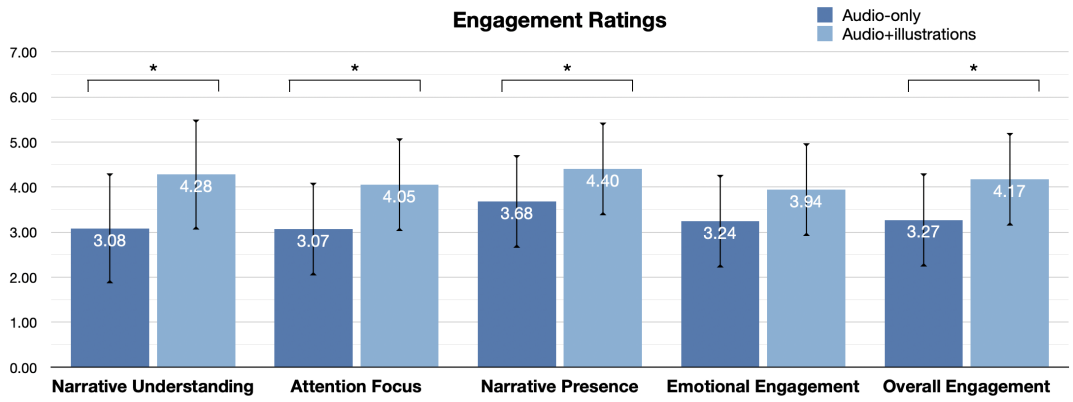

Figure 5: Mean scores for narrative understanding, attention focus, narrative presence, emotional engagement, and overall engagement in the audio+illustrations modality compared to the audio-only modality.

3.5 Discussion

Our study sought to understand the potential benefits of integrating illustrations with audiobooks in multitasking scenarios. We have organized this section in order of each hypothesis, followed by continued discussions at the end.

H1: The audio+illustrations modality leads to better content recall immediately and after 7 days when compared to the audio-only modality. Our findings support this hypothesis. Participants exposed to the audio+illustrations modality showed a 33% improvement in immediate recall and a 32.7% improvement in 7-day recall, compared to those in the audio-only group.

Conceptual pegs: We found that illustrations serve as memorable anchors for complex information. For instance, P11's recall of the "vehicle resembling a Martian rover" in The Martian track highlights the efficacy of visuals in distilling detailed and intricate narrations. While the audio description of the "Mars Ascent Vehicle" was detailed and complex in the audiobook, it was the straightforward visual representation of "the rover" that offered a concise, memorable anchor. This supports the concept that visual and verbal encodings can work synergistically to enhance recall as suggested by the Dual-Coding Theory [86]. According to this theory, when both verbal and visual information are provided, they offer two ways to encode the information and two ways to retrieve it later, which can enhance memory and learning. These results also align with the Cognitive Theory in Multimedia Learning [71], which argues for the effective utilization of both visual and textual (or auditory) modalities to enhance information processing and retention, given that the human working memory has limited capacity within modality-specific stores.

Contextual misrepresentations: Illustrations can contain rich information, sometimes conveying even more contextual detail than the audio narration alone: "There were four people in the crew (or at least that's what the illustration suggests, the audio didn't make such a claim)" (P11). While there are advantages to the additional 'code' offered by illustrations, P12's misinterpretation of the "Mars Descent Vehicle" as "balloon-like" serves as a cautionary note that while illustrations can clarify, they run the risk of misrepresentation and oversimplification. To address this, a human-AI collaborative approach is recommended where human illustrators offer the expertise and contextual understanding that AI currently lacks [25] (Section 6.6 on Ethical Concerns).

H2: The audio+illustrations modality is rated as more engaging compared to the audio-only modality. Our results robustly support H2, revealing that the audio+illustrations modality outperforms audio-only in narrative understanding, attention focus, narrative presence, and overall engagement. While emotional engagement trended in a similar direction, it narrowly missed statistical significance.

Emotions and visual storytelling: Drawing on the Cognitive Affective Theory of Learning with Media [76], emotionally engaging multimedia has the potential to boost motivation and deepen cognitive processing, resulting in enhanced learning outcomes. P5 highlighted the role of visual storytelling elements in enhancing emotional engagement and sustaining user attention [42], by describing the illustrated modality as "more immersive", and that they felt "more emotionally attached to the character".

Impact on Imagination: A notable theme that emerged was how accompanying illustrations affect the listener's capacity for imaginative interpretation. Existing research on stories suggests the important role mental imagery plays in comprehension, recall and meaning-making [37, 92]. While the absence of visuals in the audiobook format leaves important gaps for listeners to form their own mental imagery of the story, videos on the other hand present continuous visuals, providing greater immersion. Each medium rich and unique in their own right, AudioXtend is designed for multitasking and thus aims to strike a balance between both. That is, illustrations presented on OHMD add a visual dimension to audio storytelling, depicting key portions of the narration that users can glance at while going about routine tasks. The simple outlines-only

style of illustrations and low frequency of visual stimuli relative to motion pictures creates some space for imaginative interpretation.

Interestingly, P1 and P13 found that the illustrations helped them "imagine the story better," suggesting that the illustrations used in Study 1 did not overshadow their imaginative process, but instead, supported it. P12 also noted that it felt effortful to generate their own mental imagery, particularly when multitasking; the illustrations scaffolded this process, reducing their cognitive load. These pieces of feedback provide positive indications of AudioXtend's approach, and it remains essential to consider cognitive processes such as imagination and more in future designs of AudioXtend (discussed further in Section 6.1).

H3: There is no difference in primary task performance, perceived disruption to primary activity and perceived task load between the audio-only and audio+illustrations modality. Our findings indicate that H3 was partially met. While objective measures of primary task speed showed no significant modality-induced differences, suggesting that the inclusion of illustrations did not detrimentally impact task efficiency, participants' subjective ratings of their multitasking experience revealed a more nuanced account. Specifically, when asked "How smooth was your experience of executing the primary task (walking navigation or folding clothes) while simultaneously engaging with the audiobook content or viewing illustrations?" participants rated walking as a 42.4% smoother task to execute than folding laundry.

Task interference and automaticity : This may be attributed to the eyes-busy nature of the folding laundry task: P4 found it "harder to focus when folding clothes compared to when [they were] walking", similarly, P6 found themselves "easily distracted during the process of folding". This can potentially be accounted by the Multiple Resource Theory, which suggests that tasks interfere when they tap into the same cognitive resources [10, 118]. Folding laundry competes for attention with the on-screen visuals, given that both tasks utilize cognitive resources from the same verbal-pictorial dual-channel [121, 124]. In contrast, walking, particularly in familiar settings, is a more automatic task that requires fewer cognitive resources [31].

The question of primary task interference also touches upon the concept of automaticity, which refers to the ability to perform tasks without conscious thought, achieved through repeated practice and familiarity [9]. When tasks reach a level of automaticity, cognitive resources are freer, enhancing multitasking capabilities [65].

Our findings in Study 1 underscore the effectiveness of choosing "autopilot" primary tasks such as walking along familiar routes. These tasks can be prioritized for the purpose of use with AudioXtend, for they assure steadier multitasking performance through minimized cognitive load. While incorporating more complex tasks, such as building furniture or solving puzzles, could offer additional insights, our focus is on incidental learning and the everyday applicability of audiobooks. Complex and cognitively demanding tasks may therefore not accurately represent the common contexts in which audiobooks are consumed. Understandably, participants cautioned against using AudioXtend for critical tasks that pose safety risks, such as crossing a road, or tasks demanding constant visual attention. Encouragingly, even under less-than-ideal conditions, AudioXtend continued to provide an improved experience to the audiobook, e.g. P1, P2, and P4 noted that despite the visual busyness of clothes-folding, they still favored the OHMD visual augmentations for its contributions to understanding and memory retention.

Glanceability: Study 1 also revealed a critical tension between glanceability and information richness, particularly in the context of OHMDs and multitasking. Glanceability refers to users' capacity to swiftly and seamlessly extract pertinent information from a visual display without detracting from their primary task [68] and relates to immediate comprehensibility and non-obtrusiveness in immersive systems [13]. Participant feedback revealed competing desires, and the balance that needs to be struck between glanceability and information richness [11, 69, 111]. While some sought more frequent, detailed illustrations for a richer narrative experience, others advised caution, particularly when multitasking demands intensify. For instance, P6 and P7 indicated that they "prefer the illustrations to be more frequent" because "the illustrations helped [them] follow the storyline", while others believed that too many illustrations could jeopardize multitasking. P11 consequently suggested a "lower frequency during heavier multitasking". In this regard, AudioXtend markedly differs from traditional video-based learning, offering key information through selective visuals that support glanceability on OHMD in multitasking scenarios. In contrast, videos demand continuous visual attention and are likely to hinder the ability to simultaneously manage other tasks.

Further considerations emerged from the interviews, such as visual complexity. Some asked for more vivid, detailed illustrations (e.g. P11 suggested coloring for saliency: "Highlight important details per frame with colors"), while a minimalistic style appealed to other participants. P12 proposed "no color as it is distracting", and as explained by P8, "[as illustrations get more complex], you'll need to put more effort to understand...the aim is to help you follow the story, so you don't need to be too detailed." Research by Somervell et al. supports this, noting the search time-cost of high-density visuals [106]. Participants suggested other design factors such as illustration size, and multiple-frame layouts over a single fixed frame. This introduces the potential for customizable design features, allowing users to tailor parts of their own visual experience. Acting on these suggestions, we decided to implement some of these customizable features into the AudioXtend application used in the next studies.

While Study 1 has shown that the audio+illustrations modality positively impacts recall (immediate and longer-term) and engagement levels without compromising significantly on primary task efficiency especially for primary tasks that participants consider more autopilot, the variety of emerging factors calls for a more nuanced exploration of visual augmentation design factors for AudioXtend. Thus, we decided to proceed to a participatory design workshop to explore the "what, when and how" that makes visual augmentations more effective or engaging than others.

4 PARTICIPATORY DESIGN WORKSHOP

After the completion of empirical Study 1, which explored the impact of visual augmentations on memory recall, we now shift our focus to visual design— seeking better understanding of "what" makes certain illustrations more effective or engaging than others. To do so, we organized a Participatory Design Workshop, a method

well-suited to our needs given its collaborative and user-centered nature [102]. The broad and complex design space for visuals warranted a hands-on approach that would allow us to tap into the collective wisdom and creativity of those who have some experience in digital art editing and AI-image generation. By involving these individuals in the co-creation process, we aimed to arrive at a group consensus about the most effective and engaging design choices, thereby enriching the guidelines we had begun to formulate in our earlier studies.

4.1 Methodology

4.1.1 Participants. We recruited 10 participants (5 male, 3 female and 2 non-binary; Mean age = 22.7, SD = 2.67 years), all of whom have self-reported experience in digital art editing and AI-image generation. 6 were native English speakers, and the other 4 non-native speakers self-reported that they were fluent. Universally, participants had prior exposure to audio-first media forms, such as audiobooks and podcasts, which was a criteria included for meaningful participation. Each participant was compensated ≈ $\$7.5/h$ for their time. All studies were approved by the IRB of our institution, and an informed consent was obtained from every participant.

The majority (9 out of 10) described their expertise with AI-image generation tools as moderate with sufficient understanding of the tools, while one participant indicated substantial use and fluency. DALL-E, Microsoft Bing, Midjourney and Stable Diffusion were the tools they used, with the first two cited as most common. For digital art editing, Photoshop and Canva emerged as the go-to platforms. A subset of participants also had experience with other softwares, including Illustrator, Procreate, Lightroom, Sketchbook, and Vectornator. To gauge their skill level and offer a warm-up practice, we issued a trial task prior to the workshop where each participant had to use their preferred AI image generator and editing software to generate two AI-created images based on a specific audiobook sentence.

4.1.2 Workshop Venue. Designed to simulate real-world multitasking conditions, the sessions featured a variety of tasks set up in the workshop venue for participants to engage in. These included a laundry-folding station to mimic domestic chores, a tea/coffee preparation station complete with unwashed fruits to represent kitchen tasks, and open spaces where participants could walk around, simulating on-the-go conditions. Throughout the workshop (see Section 4.1.3), participants were asked to test visual designs on OHMD whilst performing daily routine tasks, i.e. rotating through the different stations (Figure 6).

4.1.3 Workshop Format. 3 workshop sessions were conducted, each featuring between 3 and 4 participants, and lasting for 2 to 2.5 hours, including self-introductions and a briefing. Supervised by a team of 3 to 4 facilitators, these participatory design workshops were structured into four segments:

(1) **Individual Listening Session (10 minutes)**: Each participant was provided with their unique audiobook track (i.e. each participant was assigned a different track), which came with our default set of visual augmentations displayed through OHMDs. Participants were instructed to listen to the audiobook while concurrently performing real-world primary tasks, such as folding laundry or preparing tea. The aim was to allow them to experience first-hand the integration of these visual aids into real-world multitasking scenarios. This session served as an orientation for their design activity in the next segment of the workshop, during which time they begin to understand what they will be designing for, and what their personal design preferences are.

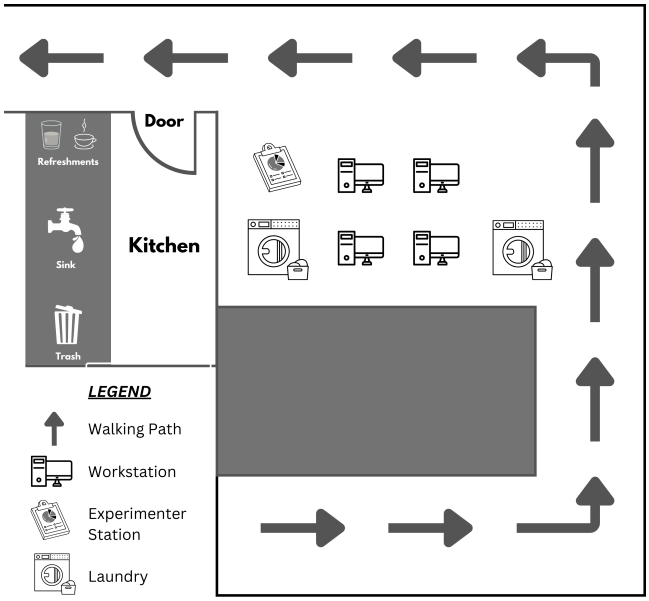

Figure 6: Floor plan of the workshop venue.

(2) **Design Session (60 minutes)**: Participants were tasked with creating illustrations to accompany a 2-minute segment of their assigned audiobook track. We provided the flexibility to use any AI tool or editing software for their creations, and advised that they create a total of 6-10 illustrations for the 2-minute track, which is a comfortable frequency range based on previous pilots. Aside from advising that black backgrounds would appear transparent on their display screens and requesting square dimensions for compatibility, we imposed no constraints on style or content. Once their illustrations were ready, they were instructed to upload them into a designated folder with marked timestamps indicating when each visual should appear during the audiobook track. These user-generated visuals were then integrated into a custom Unity application [115] on android phones for synchronization with the audiobook.
Throughout the design session, participants were also asked to maintain a design log, capturing their choices and considerations behind each visual element (Figure 7 demonstrates the workshop's workflow).

(3) **Peer Testing and Feedback (30 minutes)**: Participants tested designs created by their peers on OHMDs, while engaging in the different real-world tasks. They filled in a

peer-review survey at the end of each test, to identify the characteristics of their favorite and least favorite image for each track, as well as provide general feedback.

(4) **Group Discussion (20 minutes)**: Subsequently, they engaged in a group discussion, with emphasis on the peer-rating exercise to identify shared insights on design choices. Topics covered included but were not limited to recurring elements in high-rated designs and the feasibility of implementing certain design aspects in real-world scenarios. The discussion also touched on the adaptations needed for multitasking scenarios and garnered advice for future design work.

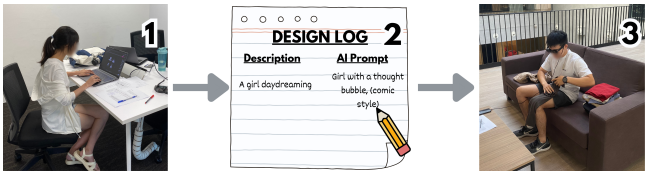

Figure 7: The participatory design workshop workflow. After the individual listening session with our default set of visual augmentations, participants (1) engaged in a hands-on design session, creating and time-stamping illustrations for a 2-minute audiobook extract. This is followed by (2) maintaining a design log to capture the description and rationale behind each visual element, as well as recording keyword searches and AI prompts used. Next is peer testing, where participants (3) evaluated each other's designs during real-world tasks, and completed a peer-review survey.

4.2 Findings and Discussion

In this section, we synthesize key insights from individual design logs, peer review, and group discussions to outline guiding principles and theoretical considerations for the future development of AudioXtend. We start first by identifying design pitfalls that could lead to disruptions during real-world tasks. Thereafter, we will explore the broader design spectrum, discussing how style and adaptive illustrations can be tailored to individual preferences.

4.2.1 *Causes of Disruption.*

Misalignment. Favored images in the peer-review were praised for their alignment with the audiobook's themes and locations, e.g. "the image of a woman matched very well with the description in the audio" (D5), and could "well supplement the audio information" (D7). On the flipside, inconsistencies or gaps in visual representation can lead to increased cognitive load, disrupting the user's ability to perform parallel tasks efficiently. A common example of this occurs when participants expect an illustration to accompany a new narrative element, but find it missing, leading to increased cognitive load.

This misalignment can manifest in other ways too, such as when the audio content diverges from its visual elements, e.g. "The tone is sad but the image is quite funny" (D11, Figure 8-(a)) and "the ship image also takes one out of the experience of the like space and stars" (D11, Figure 8-(b)). Well-crafted, aligned visuals serve to reduce cognitive strain, and improve immersion.

Figure 8: Top row: (a) Mismatched Emotions where a somber funeral scene is incongruently paired with an almost comical, floating astronaut, leading to a tonal dissonance, (b) a ship image diverging from the audiobook's narration of outer space, potentially detracting from the immersive experience, and (c) ambiguity between the black hair and the black background. Bottom row: Inconsistencies in art style between (d), (e) and (f) disrupts the viewer's attention and understanding of the scene – (d) depicts the moment when characters ate cake, (e) represents the sister and her friend in the garden, an essential narrative element but with a stylistic departure from (d), and (f) shows untouched party games, highlighting the absence of guests, albeit once again, challenging the stylistic coherence of the set as a whole.

Ambiguity. The ambiguity of the illustration itself can also add to mental load and user confusion. Ambiguous visuals, unclear symbols, or vague scenarios contribute to a lack of understanding, thus imposing unnecessary cognitive burdens. As some participants shared, "the idea was a bit too abstracted for this one, and it ends up being more confusing than helpful. It could've been either removed (and subsumed by a more suitable frame) or made more concrete" (D8). Poor color choices also affected immediate recognizability, e.g. "background could be more distinct and relevant to contrast with the black coloured hair" (Figure 8-(c), D8), while in appropriate cases, D4 noted that "high-contrast colors are easy to notice" and can remove ambiguity.

Consistency. Upholding a uniform style for the illustration set emerged as a significant factor. Particularly when using AI-generated images, maintaining a consistent style can be challenging due to the inherent variability in the visuals produced. As one D11 noted (Figure 8-(d)(e)(f)), "the jump in art style made it difficult [for viewers] to process. The image was just vastly different in tone and content from the previous pictures. The purpose of the image was a bit lost on me, and I also lost focus on the task at hand as I had to pause and think..."

Complexity. Finally, the complexity of the illustration plays a significant role: Overly busy or intricate visuals can become distractors themselves, diverting attention from the primary task. For instance, D7 rated one image lowly, because it "was quite busy and had a lot going on… harder to multitask while processing this particular image", and suggested that "perhaps a simpler design with less things could help."

4.2.2 *Design Spectrum.*

Simplicity/Immersion Trade-off. When faced with cognitively demanding tasks, participants expressed a preference for simplified and less obtrusive visual elements. Aspects such as reduced detail and increased transparency were commonly cited as ways to achieve this goal. This inclination towards transparency is notable for OHMDs, where digital elements are overlaid onto the user's real-world view. Enhanced transparency minimizes the obtrusiveness of these overlays, facilitating a more seamless multitasking experience via the see-through display [88]. As participant D5 echoed, there is a need for "less detail and a more transparent background."

Additionally, there was a consensus on focusing more on key objects within the visuals. Images in the peer review that were well-received generally shared the characteristic of having "a clear and strong subject center" (D8). In a similar vein, easily recognizable subjects that are conceptually easier to grasp are essential for multitasking where attention switching is prevalent.

To achieve greater visual simplicity on OHMDs, participants indicated a preference for visuals with less depth perception (refer to Figure 9 (a) for an illustration with high depth perception). Although extensive research discusses depth perception in traditional displays [57], literature specific to OHMDs and multitasking is limited. Our findings suggest that reduced visual depth is favored in OHMDs, especially in the context of multitasking, to facilitate easier information processing.

Considerations around Color. In continuation from the notion of having strong subject focus and recognizability, many participants had also included colors (in contrast to grayscale illustrations) in their designs, and found that high color contrasts were especially memorable. According to D4 and D8 respectively, "[the high-contrast colors] easily got my attention", and "makes it easy to focus and makes the subjects in each picture immediately recognizable". Colors can heighten emotional engagement and aesthetic appeal in the visuals, reinforcing Don Norman's Emotional Design theories [81], which posit that aesthetics have a profound impact on the user's experience and emotional engagement. Interestingly, the Aesthetic-Usability Effect also states that users are more likely to tolerate design flaws if the aesthetic elements are pleasing [56]. It is certainly the case that participants found well-applied colors to be "informative" (D10), however, there is a need for a balanced approach as colors can sometimes serve as a distraction. D1 shared that "while walking, [they] preferred not to look at the images too much so that [they] could see where [they] were going… the coloured images were distracting." Many participants concurred that black and white images were less distracting and allowed for more seamless multitasking.

Notably, D5's approach to using colored outlines without color fills offers some compromise between the aesthetic appeal of colors and demand for reduced cognitive load during multitasking activities. This approach, as depicted in Figure 9 (c) maintains a high degree of transparency, minimizing the visual obstruction in the OHMD's see-through display.

Table 1: Overview of key aspects of favored and unfavorable elements highlighted in the peer review.

Design Spectrum	Favored Elements	Unfavorable Elements
Content Alignment	Well-aligned images in terms of timing and thematic content can supplement the auditory information.	Inconsistencies in the audio-visual depiction of characters tone, or settings were cited as a downside.
Clarity and Focus	Well-received images generally had a clear and strong subject center, aiding in audience focus. Easily recognizable subjects are essential for multitasking where attention switching is prevalent.	Poor color contrast affects immediate recognizability, and black visual elements can blend into the background and therefore end up camouflaged.
Detail and Style	Images with unique styles and high color contrasts were especially memorable and captured participants' attention, However, extensive use of color might be distracting.	Images that are too abstract lead to confusion on what the objects shown are about. It is also important to fine-tune details, with issues such as poor cropping causing misinterpretation.
Simplicity	Simplified, single-layered images were easier on the eyes. This includes incorporating less visual depth and a plain background.	Overloading visuals with too much information was distracting for the viewers, affecting their ability to effectively multitask.
Frequency	A dynamic approach to determining the frequency of illustrations was favored. This approach adapts to the pace of the narration and minimizes visual distractions during static scenes, such as dialogues between characters.	A static frequency rate for illustrations can lead to unnecessary visual changes that cause distractions from the primary task.

To nuance this discussion, participants expressed that colors should adapt to the 1. primary task busyness as well as 2. the context and storyline. D1 felt that "if it's a simple activity, [they] would prefer colors", and vice-versa. As an example of how a participant made color or stylistic choices based on the mood and storyline, D11 shared in hindsight that they most likely went with the grayscale sketch style, to effectively convey the "ambiguity" of the recounted scene in the story. D3 expressed the downside to selecting colors that are not fitting to the theme of the story, e.g. "the black and white doesn't convey 'the morning' " (D3).

Figure 9: Examples of designs with (a) high depth perception (b) full colors, which may be more distracting for users, compared to (c) a design with colored outlines and high transparency, offering a balance between aesthetic appeal and reduced cognitive demands.

This reminder to align illustrations to the content and context comes as no surprise – existing literature has long emphasized the importance of adapting stylistic elements to enhance the user experience. For example, Tufte's theory on information design stresses the importance of context in determining effective design [113], while Mayer's multimedia learning theory suggests that style of each medium can affect this outcome of learning [71]. This extends into the point regarding timing and frequency.

Adaptive Timing and Frequency. In Study 2, the assumption was that a fixed frequency of illustrations would optimize the user experience. Contrary to this, D8 was one of the participants who advocated for a dynamic approach, where the frequency of illustrations is aligned with the narrative's pace, making it adaptive to the story's momentum. This perspective aligns with the dynamic pacing found in interactive narratives [78] and underscores the need for a more fluid approach to visual augmentation. In fact, participants suggested that a strong narration with moving plot points should result in a higher frequency of visuals to capture the dynamic story elements: "If you're describing new objects, new settings, you could speed up the frequency" (D5). Conversely, if the setting, characters, and action in the narration remain static, such as in a dialogue between two characters, the frequency of illustrations should be minimized. As D8 shared, "[when] person one and person two are having a conversation, you don't need to flip back and forth all the time."

As a caveat, it is crucial to consider an upper bound for frequency, as excessive changes can distract users: "High frequency is very distracting when images change too often" (D3). On this point, participants again echoed the strong need for adaptive illustrations that are responsive to both the complexity of the scene and the cognitive demand of the primary task.

Usage of Generative AI. Generative AI tools can be used as a starting point for creating and testing the style of the illustrations. The objective is to streamline the creative process by leveraging AI as a collaborative tool rather than as a substitute for human creativity [43]. As shared by the participants, the common factors included in the prompts for creating the illustrations are "simple sketch" "black background" (D1, D4, D5, D6, D7, D11) and phrases describing view angles (D7, D8). Other interesting prompts that produced illustrations seen as favorable by other participants are "light color streaks" (D5). Notably, many participants highlighted the inconsistency of style between the consecutive illustrations as an inevitable result of AI-generated imageries. To improve the style consistency, several directions can be considered for the future researchers: 1. involve professional illustrators in the loop to edit illustrations and ensure style consistency, 2. use more advanced AI algorithms to improve style consistency. For further reading, Appendix A details a potential LLM-assisted workflow to facilitate the generation of narrative illustrations.

5 FIELD STUDY

In Study 1, we conducted lab-based studies on visual augmentations. While it provided crucial insights on how users engage with AudioXtend under controlled conditions, it does not fully capture the multifaceted nature of real-world use [22]. Thus, the aim of this field study is to delve deeper into user behavior and experiences while they use AudioXtend with their routine daily activities.

Conducted over a span of three days, this take-home study allowed for a more extended and realistic user engagement period, where participants had the choice to use it in any usual setting of their choice — be it at home, during their commute, or in other on-the-go scenarios. Participants were exposed to one hour of audiobook content (cumulatively spread out over the whole study). The study seeks to uncover the specific challenges users encounter and the preferences they exhibit when interacting with AudioXtend, all within the context of their natural, everyday environments.

5.1 Participants

6 participants (3 Females, 3 Males) from the university community. Their mean age was 25.3 (SD=4.04) years. All were fluent in English, and spoke it either as their first or second language. All had prior

experience listening to audio-first content (e.g. podcasts, audiobooks), and 2 out of 6 participants had used smart glasses before. Each participant was compensated ≈ $\$7.5/h$ for their time, i.e. a total of ≈ \$11.25 for their complete 1.5h participation, including their briefing, AudioXtend use, and final interview. All studies were approved by the IRB of our institution, and an informed consent was obtained from every participant.

5.2 Apparatus and Materials

Each participant was loaned a pair of OHMD and wireless headphones (same as that used in the previous studies), and an android phone with the AudioXtend Field Study application, implemented using Unity. The application contains two audiobook tracks, Dogs of Riga and The Ocean at the End of the Lane, both of which were tracks used in Study 1 and 2, albeit now extended to 30-minutes each.

To make the application more user-centric, design elements informed by the previous study and design workshops were incorporated into the illustrations. These elements were characterized by simplicity, using simple white outlines against a transparent backdrop, and standardized content alignment to ensure greater consistency with the narration. Additionally, a midpoint frequency of 3 illustrations per minute was set to avoid information overload. Customizable settings for a multi-pane "Timeline" view and Scale (see Figure 10 for examples; recommended from Study 1) were also integrated, enabling users to further personalize their experience.

5.3 Procedure and Data Collection

Participants were asked to choose three consecutive days to partake in the study, during which they had to complete two 30-minute audioXtend sessions. They were free to select their own session timings but were encouraged to do so during different daily activities of their choice to capture diverse contexts (refer to Figure 11 for examples). After each session, participants filled out a survey administered via Google Forms on their own device to answer multiple open-ended questions:

- Describe all the real-world tasks you were engaged in during this audiobook session.
- What made you think that this was a good time to experience the audiobook?
- What in-app settings did you prefer during your session and why?
- Were there any issues or frustrations?
- Were there moments you found particularly enjoyable?
- How was this session different from the previous one?

As part of the survey, participants were also asked to rate their experience on a 7-point Likert scale with questions focusing on primary task disruption, story engagement, content recall, emotional or atmospheric enhancement and overall experience. These areas of assessment allowed us to evaluate AudioXtend's impact in ecologically valid settings.

To supplement the data, participants were also asked to have photos or videos taken of them during the sessions (Figure 11). Upon completion of the two audiobook sessions, a 15-minute open-ended Zoom interview was conducted to gather additional insights. The results are discussed thematically in the subsequent section.

Figure 10: Top row: A multi-pane timeline features a vertical strip of illustrations, with a main viewing pane, a top pane previewing upcoming content, and past content moving out of view via the bottom. The strip moves downwards by one pane when illustrations change according to the narration. Bottom row: The customizable Scale option allows users to adjust illustration sizes.

5.4 Results and Discussion

Our field study data reveals key insights into both the strengths and areas for improvement for AudioXtend in real-world scenarios, providing crucial guidance for enhancing its future usability and user experience.

5.4.1 Immersion after acclimatization. 5 out of the 6 participants needed up to 30 minutes to get used to AudioXtend, while one participant felt immediately comfortable with its use (F4: "It felt natural"). Most of the initial discomfort pertained to hardware issues like the OHMD lenses (discussed further later), e.g. "The smart glasses lens were tinted, it was hard to differentiate colours when I was folding my clothes" (F5).

As participants settled into the system, all of them expressed appreciation for the system's capabilities, which allowed them to engage with audiobooks in a more immersive manner. Echoing the sentiments of participants from Study 1 and the design workshop, F1 pointed out that the visual augmentations helped them "understand the content better" and especially in real-world settings where environmental distractions abound, visuals helped them stay "focused" when they "miss out on the plot [from the audio narration]". The visuals not only captivated their attention but also served as a safety net, capturing the essence of the storyline when distractions arose (F2: "When I spaced out and missed out on the audio, fortunately the images were still there for me to capture the

Figure 11: Participants experiencing AudioXtend while performing daily activities at home and in public, such as 1) doing laundry, 2) organizing study materials, 3) cooking food, 4) washing dishes, 5) shopping at a supermarket aisle, and 6) taking a walk.

main gist"). He contrasted this with auditory information, which is much more fleeting: "With audio you only listen to the words once and it passes, but the visuals stay for a while".

F5 also pointed out that the visuals allowed them to "maintain concentration consistently [throughout the 30 min session]", which is less likely in an audio-only implementation. Especially while engaged in low-attention tasks, participants reported immersing deeply in the narrative, emotionally connecting with story characters. In F2's words: "I felt engaged in the story, so much so that I had a reaction when…the first kitten was run over." Participants not only reported improved focus but also felt they absorbed content efficiently and naturally. F6 likened the experience to "reading a book with illustrations," underscoring the naturalness of the interaction.

They felt that the overall experience served the purpose of immersing themselves in the narratives, improving their concentration even during mundane tasks, ultimately leading to a positive user experience. F1 and F2 agreed that the existence of the visuals helped them capture the gist of the storyline even when they get distracted by their environment which they would have missed if there were no visuals.

5.4.2 Chosen moments of use and effect on daily tasks. Using AudioXtend in everyday contexts, 2 participants (F5 and F6) chose to listen to the audiobooks when they were out and about, either commuting or shopping (Figure 11-5, 6). F6 found that AudioXtend enhanced her commuting experience by providing visuals that "acted as a point of focus, but not too intrusive or imposing so that I can glance at it from time to time to stay in tune with the story". Both of them chose to use AudioXtend on-the-go as they felt that the audiobook was a good substitute for music. As F5 shared, "listening to a story seemed more engaging than music and informational podcasts".

Meanwhile, the 4 other participants chose to use AudioXtend predominantly in domestic spaces (Figure 11-1, 2, 3 & 4), such as while eating, cooking, exercising, and completing chores. F1 noted that AudioXtend made exercising more endurable: "During my plank exercise, the audiobook managed to capture some of my attention, making me feel less tired compared to my previous workouts." These participants often chose to listen to the audiobooks whilst tending to domestic tasks given their triviality and repetitiveness. F3 expressed: "Doing laundry is a repetitive and mindless task that does not engage my executive functions much. I thought I would be able to engage more with the audiobook while doing this." This same reason prompted participants to "opt for an audiobook as a slight entertainment besides these simple tasks" (F5).

5.4.3 Points of discomfort in real-world contexts. Key safety issues emerged, primarily linked to the hardware of the system. Several of these issues have been found in prior research using smart glasses such as reflections on the lenses (F6: "my vision was affected by the reflection"), altered color perceptions ("difficult to pick the fruit that I want…lens and images affected the color I see"), and reduced visual contrast in bright sunlight (F5: "white/light colors did not have much contrast against the sunlight). In addition, a majority of the study participants' expressed discomfort or self-consciousness stemming from perceived public perception while wearing smart glasses, a common predicament when using wearable devices in public [17]. It is worth noting that many of these challenges may be attributed to the current generation of smart glasses, which are often bulky and conspicuous. However, as smart glasses evolve to resemble conventional eyewear more closely, we anticipate that both safety and social acceptance will improve significantly. This evolution would reduce the visual obtrusiveness and ergonomic discomfort, thereby making systems like AudioXtend increasingly feasible and user-friendly in everyday scenarios.

5.4.4 Potential of AudioXtend. Notwithstanding the pain points expressed by participants, their overall response signaled high satisfaction (the mean rating was a robust 5.0 out of 6) and strong interest in continuing the use of AudioXtend beyond this study. When prompted to reflect on what would motivate them to use AudioXtend on an everyday basis for the purpose of incidental learning, two main areas for improvement emerged: Customization features and hardware design.

For instance, F2 and F3 asked for full playback options such as the rewind function to revisit specific audiobook sections. F4 and F6 expressed their preference for coloured illustrations, with F6 mentioning in addition their fondness for "comic or manga style". Moreover, F5 hoped for greater "freedom of choosing the placement of the illustrations" stating that optimal positioning of the illustrations could vary depending on his task at hand. These suggestions offer insights for refining future versions of AudioXtend.

Encouragingly, 5 out of 6 participants expressed keen interest in experiencing AudioXtend as a continued user. This was as discussed, contingent on the OHMD being more streamlined, lightweight and stylish, particularly for out-of-home usage. On the other hand, F3 highlights an entirely different motivation for using audiobooks, and thus prefers to discontinue using AudioXtend. F3's perspective was that he preferred to use audiobooks as a form of atmospheric, mood-based stimulation rather than information-driven experience. For this reason, visual augmentations do not add to his use. It is important to note that our primary goal is to enhance incidental learning experiences (discussed in the next section), and

consequently do not recommend AudioXtend for other purposes of audiobook consumption.

6 DESIGN RECOMMENDATIONS AND FUTURE WORK

In this section, we integrate key design principles informed by our workshops and studies, providing future directions for designers and researchers interested in building on the AudioXtend platform.

6.1 Design Considerations for Enhancing Incidental Learning in Audiobooks

There are a variety of motivations to audiobook-listening, including language learning, improving literacy and relaxation [73]. While our focus is on enhancing informal and incidental learning in multitasking contexts, we recommend a holistic perspective on design for AudioXtend. The functional aspects of visual design include:

- **Labeling Illustrations:** Annotations or labels of characters and locations have been shown to help users interpret and integrate information provided by the AI-generated illustrations. Future designs could consider offering the option to toggle labels on or off, allowing users to customize their interaction with the technology according to their personal preferences and needs.
- **Consistency of Story Elements:** A consistent alignment of visual and audio elements is crucial for bolstering memory and creating a cohesive and immersive experience that supports the listener's ongoing engagement.
- **Clarity Over Complexity:** We recommend the use of transparent outline images and minimal color usage. This minimizes cognitive load, facilitating a seamless experience without causing distraction or confusion.

Beyond mere functionality, AudioXtend can utilize AI-generated visuals as a method closely aligned with our cognitive processes, rather than just as an additive element. Forgetting [8] is an example of a cognitive process that filters sensory information, enabling more effective thinking and action. This concept highlights the importance of filtering sensory information for effective thinking and action. We propose integrating selectively timed illustrations in AudioXtend that can fade or disappear after a short duration, encouraging users to focus on the story's essential elements without being overwhelmed by persistent visual stimuli. This design choice would mimic natural cognitive rhythms [49], helping listeners focus on the essential elements of the story without the cognitive load of persistent visual stimuli. Mind-wandering [5] is another cognitive process that can be better designed to align with the natural rhythm of memory and attention [30]. To support this common function of creative incubation, future work can explore ideal moments of 'blank space' or minimal visuals. These intervals can be strategically placed at points in the narrative where reflection or imagination would be most beneficial, enhancing the listener's engagement with the story through opportunities for mind-wandering.

We recognize that at the core of AudioXtend's approach lies the implicit pursuit to maximize human efficiency and productivity in daily life. As highlighted by Lin and Lindtner [63], design and computing methodologies often contribute to fostering "good" and "positive" feelings associated with productivity and progress, yet this emphasis often occurs without critical reflection on the underlying social and ethical implications. This oversight, which AudioXtend shares, risks perpetuating harmful ideologies and reinforcing existing power structures. Particularly, it may contribute to silencing and normalizing the violence experienced by individuals who are not "yet" seen as productive, useful, able, or innovative. It is thus essential to thoroughly consider the productivity-centric value systems and socio-technical landscapes within which AudioXtend and other similar approaches operate. This allows us to rightfully explore and mitigate potential harms associated with them.

More broadly, embracing the fourth wave HCI approach involves recognizing the growing entanglement of technologies with our bodies and daily lives, as well as the increased complexity of interpreting empirical observations [33]. As we move towards future AudioXtend designs, it is crucial to adopt a more integrative approach that not only acknowledges the different components of the socio-technical ecosystem, but also actively engage with real-world issues around politics, values and ethics [3].

6.2 Familiarizing Users with AudioXtend and Guidelines for Safe Usage

Our field studies indicate that participants generally felt more at ease with the smart glasses during their second than first session. This underscores the importance of a well-designed introduction to familiarize users with the technology and its capabilities. It is crucial that users are made aware of the recommended contexts for utilizing AudioXtend, differentiating this from the experience of traditional audio-only formats.

Here, task suitability is key. AudioXtend is best used during "autopilot" tasks, such as doing dishes, walking, or light exercise, as they allow for incidental learning without straining cognitive resources. On the other hand, tasks that require complex problem-solving, critical decision-making or constant situational awareness should be avoided, such as when writing an email, navigating through a crowded place or crossing the road. To further enhance the user experience and safety, we propose that the system should have the capability to adapt to the user's environment and current state.

6.3 Building an Adaptive User-Context Model

One critical avenue for extending the capabilities of AudioXtend is to develop an adaptive model that takes into account various situational factors:

- **Frequency:** Dynamically adjust illustrations based on user interest and context. This could mean hiding illustrations or reducing frequency when a user is, for example, crossing the street or focused on another task that requires their full attention. More fundamentally, future works should first explore the optimal frequency of illustrations in relation to key story events. This may involve determining whether illustrations should be triggered by specific narrative cues such as major plot developments, character introductions, or

changes in setting. The aim would be to enhance users' experience, while maintaining a balance that avoids cognitive overload.
- **Modality:** Switching between different forms of illustrations, text, or even haptic feedback [52] based on the context of where the user is at, or what they are doing.
- **Placement and Scale:** Understanding the user's environment and adapting the location and size of illustrations accordingly. For instance, smaller images when in crowded spaces or larger, and more prominent images for relaxed settings.

Leveraging methodologies developed in prior research such as eye-tracking [54, 55, 60], we can collect insights into how these adaptations affect cognitive load and iteratively refine our adaptive model to offer a more seamless and safer user experience.

In addition, the idea of spatially anchored visual elements was requested by many of our participants. This is an interesting area of expansion given the envisioned use of AudioXtend in everyday spaces. Future work can build on research like the transformation of 2D graph data into 3D AR spaces (e.g. [103]) and the development of location-aware AR experiences in everyday settings (e.g. [75]). Integrated digital objects into the user's three-dimensional spatial context is a promising area for enhancing immersion.

6.4 Allowing Customizable Interface Configurations

One of the standout findings from our field studies was the substantial variation in user preferences and needs. This variation reflects the diversity in habits and learning styles among participants. Here are two examples of user-personalizable features that can be built upon:

- **Style of Illustrations:** Each track can offer different styles of visual augmentations to choose from, ranging from realistic images to more abstract illustrations.
- **Inclusive Designs:** The availability of customization options naturally leads to a more inclusive design. For example, users with color vision deficiencies could be given the option to choose from various color schemes that are more accessible to them [120]. Additionally, users with specific learning needs, such as dyslexia, might find certain types of illustrations or text overlays more comprehensible than others [89, 90].

6.5 Accelerating the Design Process with Thematically Consistent Image Generation

Our initial focus on fiction genres with strong narrative arcs and well-developed characters has proven compatible with the AudioXtend technology, showing promising levels of engagement and recall. However, the underlying aim is to make this technology versatile enough to enrich a broader range of audio-first content. This includes but is not limited to, podcasts, news broadcasts, and audio guides for museums and tours. The objective is to supplement any narrative or informational audio content with glanceable, meaningful visual aids displayed on OHMD.

Our study employed a mixed-method approach for generating illustrations, as detailed in our LLM-Assisted Workflow (Appendix A). Acknowledging current limitations in AI-generated images, an area previously explored in works focused on the consistency of AI-generated story scenes [40, 61], we envision that advancements in machine learning will enable users to curate custom datasets for training models. This would facilitate the diversification of audio-first content across various genres. This future development will require not only advancements in machine learning algorithms but also user-friendly tools for dataset curation and input.

Furthermore, we envision that the current AudioXtend creation workflow can be automated further, potentially allowing for on-the-fly image generation. However, it is important to note that manual curation and generation were essential in our study to ensure the quality, thematic consistency, and emotional resonance of the illustrations. AI-generated art has its current limitations in capturing the nuances and context-specific details required for our study's objectives. More prominently, in light of ethical considerations tied to AI-generated content, we recommend that manual curation remains a component of the image generation process to ensure that ethical standards are maintained.

6.6 Ethical Concerns about AI-generated Art

Emerging tools for AI-generated art offer exciting possibilities for creative expression but also raise significant ethical concerns, particularly in the context of applications like AudioXtend, where AI-generated art is paired with audio content. One prominent ethical challenge is the issue of copyright and authorship. AI-generated art often incorporates contributions from various sources, which can be especially problematic when the AI model is trained on copyrighted images [97]. This raises questions about authorship — whether the creator is the machine, the AI programmer, or the original artists whose work influenced the AI's learning process. The technology behind widely-used LLMs such as GPT-4 erases the individual contributions of artists whose works have been incorporated into the training sets. This disregards the creative efforts and intellectual property rights of the artists, and as such, perpetuates a system where consent, inclusion, and power in the creative process are undermined. AudioXtend, like many other applications leveraging AI-generated content, benefits from the speed and ease of current models, albeit at the expense of ethical considerations regarding the treatment of artists and their creative output. By supporting efforts to recognize and compensate artists for their contributions, we can promote a more equitable and respectful ecosystem for creative collaboration.

Moreover, determining responsibility for inappropriate or infringing AI-generated content presents a significant challenge [114]. To address these concerns, understanding the origins and processing of data used in AI-generated art, known as data provenance, is crucial [35, 116]. Generative AI in art, as explored by Srinivasan et al. [107], must also address stereotype amplification. The propagation of unrealistic body ideals is a known risk rather than a mere potential harm of AudioXtend, as it may contribute to reinforcing societal standards and expectations regarding physical appearance. This highlights the importance of scrutinizing the impact of AI-generated visuals on body image perceptions and promoting body

positivity in media representations. This extends also to the reinforcement of harmful biases related to gender, race, and other social constructs, necessitating educational and interdisciplinary efforts to mitigate these issues [108].

Eshraghian's work [28] outlines some important principles that may guide all stakeholders. Firstly, programmers should diligently document the creative and technical processes, use appropriate software licensing when sharing code, and make informed decisions regarding user and trainer input boundaries. Secondly, trainers should record dataset catalogs and training processes with a focus on originality indicators, while users keep records of program runs with user-based inputs to manifest selectivity and curation. Lastly, contributors should ensure that AI-generated work respects the rights of others and avoids copyright infringement. These guiding principles should certainly inform the integration of AI-assisted approaches like AudioXtend, as we strive to evolve in response to the dynamic advancements in AI technology.

7 LIMITATIONS

Narrative and Task Diversity. Our initial framework for AudioXtend has been constrained by a narrow focus on specific audio narratives and a limited set of everyday tasks explored. Future iterations should aim to be versatile, tailoring the experience for a truly broad range of narrative genres and real-world activities, perhaps by utilizing the taxonomy of Activities of Daily Living [27] for a more structured approach.

Environmental Limitations. Our research was conducted in a controlled indoor setting, eliminating variables such as fluctuating weather and natural lighting that could impact the OHMD's visibility and performance, as commonly observed in prior OHMD research [36]. While this approach provides a conducive environment for focusing on narrative content, it limits the applicability of our findings to outdoor or complex settings where lighting and safety are factors.

Demographic Constraints. Our study's participant pool consisted mainly of younger adults from an academic setting, potentially skewing the results. While our findings are significant within this demographic, they may not be generalizable across diverse age groups or socio-cultural backgrounds. Further studies with expanded demographic representation could substantiate the applicability and robustness of AudioXtend [18].

Accessibility and Inclusivity. While our design principles aim to be universally applicable, we have not yet fully addressed how AudioXtend could be designed to be inclusive for users with diverse cognitive abilities [23], or variations in sensory preferences and learning styles [32]. This is a critical avenue for future work.

Novelty. The scope of our study did not extend beyond 3 days, leaving unanswered questions about how the novelty factor might affect sustained user engagement and incidental learning benefits. The novelty effect, which influences user behavior, perception, and task performance, usually occurs when users experience heightened interest during their initial interactions with the new technology [20]. This is particularly the case for users who are not accustomed to AR interfaces. To minimize novelty effects, Bach et al. [4] implemented a long-term training condition to examine the impact of familiarity on task performance, revealing that training can result in additional improvements. Tran et al. [74] also emphasized the importance of including experienced participants or conducting familiarization sessions to help users become more comfortable with the technology before the study begins. While participants in Study 1 were introduced to a sample audiobook on smart glasses during their briefing, this could have been lengthened to ensure greater familiarization. Additionally, while our 3-day field study allowed participants more time to acquaint themselves with AudioXtend, future research could benefit from an extended longitudinal study to thoroughly examine the impact of factors like the novelty effect and technology fatigue [26].

8 CONCLUSION

In this research, we have introduced AudioXtend, a technique for augmenting audiobook experiences with glanceable, AI-generated visuals via OHMDs. Designed to complement everyday activities — ranging from exercising to commuting — we leverage assisted reality technologies to enrich audio-first content. By doing so, visuals are positioned as a supportive, secondary content medium for incidental learning. Our multi-phase studies, including an initial empirical study, participatory design sessions, and a 3-day take-home field study, have provided comprehensive insights into immediate and 7-day content recall, narrative engagement, primary task efficiency and user experience. Notably, the integration of glanceable visual augmentations led to significant improvements without compromising primary task performance.

While we have demonstrated the potential of AudioXtend as a viable interface, we recognize that this work represents only one dimension in a broader design space. Future iterations of AudioXtend will build towards an adaptive model to understand the needs of everyday users and explore additional features for effective audio-first content delivery. As head-worn wearable devices continue to gain traction in everyday life, AudioXtend is a step towards designing multi-modal experiences that cater to the evolving information needs of the modern user. We hope as much to celebrate the innovations of digital technology as we do the craft of storytelling and joy of learning.

ACKNOWLEDGMENTS

We are thankful for the support that this project has received from multiple sources. It is funded by the National Research Foundation, Singapore, under two programs: the AI Singapore Programme (AISG Award No: AISG2-RP-2020-016) and the Campus for Research Excellence and Technological Enterprise (CREATE) as part of the DesCartes programme. The Ministry of Education, Singapore, has also contributed through its MOE Academic Research Fund Tier 2 programme (MOE-T2EP20221-0010). Additionally, the CityU Start-up Grant has provided partial support. We extend our gratitude to Muhammad Fahim Tajwar for his assistance with the design of figures, Yu-Rou Lin, and Synteraction Lab members for their valuable feedback in different stages of our research.

REFERENCES

[1] [n. d.]. *OpenAI GPT-4.* https://openai.com/gpt-4

[2] Cynthia Adams, Malcolm C. Smith, Monisha Pasupathi, and Loretta Vitolo. 2002. Social Context Effects on Story Recall in Older and Younger Women: Does the Listener Make a Difference? 57, 1 (2002), P28–P40. https://doi.org/10.1093/geronb/57.1.P28

[3] Simone Ashby, Julian Hanna, Sónia Matos, Callum Nash, and Alexis Faria. 2019. Fourth-Wave HCI Meets the 21st Century Manifesto. In *Proceedings of the Halfway to the Future Symposium 2019*. Association for Computing Machinery, New York, NY, USA, Article 23, 11 pages. https://doi.org/10.1145/3363384.3363467

[4] Benjamin Bach, Ronell Sicat, Johanna Beyer, Maxime Cordeil, and Hanspeter Pfister. 2018. The Hologram in My Hand: How Effective is Interactive Exploration of 3D Visualizations in Immersive Tangible Augmented Reality? 24, 1 (2018), 457–467. https://doi.org/10.1109/TVCG.2017.2745941 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[5] Benjamin Baird, Jonathan Smallwood, Michael D. Mrazek, Julia W. Y. Kam, Michael S. Franklin, and Jonathan W. Schooler. 2012. Inspired by distraction: mind wandering facilitates creative incubation. 23, 10 (2012), 1117–1122. https://doi.org/10.1177/0956797612446024

[6] Mieke Bal and Christine Van Boheemen. 2009. *Narratology: Introduction to the theory of narrative*. University of Toronto Press.

[7] Ugo Ballenghein, Johanna K. Kaakinen, Geoffrey Tissier, and Thierry Baccino. 2023. Fluctuation in cognitive engagement during listening and reading of erotica and horror stories. 37, 5 (2023), 874–890. https://doi.org/10.1080/02699931.2023.2215974 Publisher: Routledge _eprint: https://doi.org/10.1080/02699931.2023.2215974.

[8] Liam J. Bannon. 2006. Forgetting as a feature, not a bug: the dualityof memory and implications for ubiquitous computing. 2, 1 (2006), 3–15. https://doi.org/10.1080/15710880600608230 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/15710880600608230.

[9] John A Bargh. 2014. The four horsemen of automaticity: Awareness, intention, efficiency, and control in social cognition. In *Handbook of social cognition*. Psychology Press, 1–40.

[10] Michael D. Basil. 2012. *Multiple Resource Theory*. Springer US, Boston, MA, 2384–2385. https://doi.org/10.1007/978-1-4419-1428-6_25

[11] Tanja Blascheck, Lonni Besançon, Anastasia Bezerianos, Bongshin Lee, and Petra Isenberg. 2019. Glanceable Visualization: Studies of Data Comparison Performance on Smartwatches. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 630–640. https://doi.org/10.1109/TVCG.2018.2865142

[12] Alexis Boutilier. 2023. Audiobook usage increased by 70% in 2022. https://goodereader.com/blog/audiobooks/audiobook-usage-increased-by-70-in-2022

[13] Doug A. Bowman and Ryan P. McMahan. 2007. Virtual Reality: How Much Immersion Is Enough? *Computer* 40, 7 (2007), 36–43. https://doi.org/10.1109/MC.2007.257

[14] Jerome Bruner. 1986. *Actual minds, possible worlds*. Harvard University Press. Pages: xi, 201.

[15] Rick Busselle and Helena Bilandzic. 2009. Measuring narrative engagement. *Media psychology* 12, 4 (2009), 321–347.

[16] JJ Cadiz, Gina Venolia, Gavin Jancke, and Anoop Gupta. 2001. *Sideshow: Providing Peripheral Awareness of Important Information*. Technical Report MSR-TR-2001-83. https://www.microsoft.com/en-us/research/publication/sideshow-providing-peripheral-awareness-of-important-information/

[17] Runze Cai, Nuwan Nanayakkarawasam Peru Kandage Janaka, Shengdong Zhao, and Minghui Sun. 2023. ParaGlassMenu: Towards Social-Friendly Subtle Interactions in Conversations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 721, 21 pages. https://doi.org/10.1145/3544548.3581065

[18] Kelly Caine. 2016. Local Standards for Sample Size at CHI. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 981–992. https://doi.org/10.1145/2858036.2858498

[19] Russell N Carney and Joel R Levin. 2002. Pictorial illustrations still improve students' learning from text. *Educational psychology review* 14 (2002), 5–26.

[20] Wesley P. Chan, Geoffrey Hanks, Maram Sakr, Haomiao Zhang, Tiger Zuo, H. F. Machiel van der Loos, and Elizabeth Croft. 2022. Design and Evaluation of an Augmented Reality Head-mounted Display Interface for Human Robot Teams Collaborating in Physically Shared Manufacturing Tasks. 11, 3 (2022), 31:1–31:19. https://doi.org/10.1145/3524082

[21] Soon Hau Chua, Simon T. Perrault, Denys J. C. Matthies, and Shengdong Zhao. 2016. Positioning Glass: Investigating Display Positions of Monocular Optical See-Through Head-Mounted Display. In *Proceedings of the Fourth International Symposium on Chinese CHI* (San Jose, USA) *(ChineseCHI 2016)*. Association for Computing Machinery, New York, NY, USA, Article 1, 6 pages. https://doi.org/10.1145/2948708.2948713

[22] Catherine Courage and Kathy Baxter. 2005. *Understanding your users: A practical guide to user requirements methods, tools, and techniques*. Gulf Professional Publishing.

[23] Chris Creed, Maadh Al-Kalbani, Arthur Theil, Sayan Sarcar, and Ian Williams. 2023. Inclusive AR/VR: accessibility barriers for immersive technologies. *Universal Access in the Information Society* (2023), 1–15.

[24] David B Daniel and William Douglas Woody. 2010. They hear, but do not listen: Retention for podcasted material in a classroom context. *Teaching of Psychology* 37, 3 (2010), 199–203.

[25] Thomas H Davenport and Rajeev Ronanki. 2018. HBR: Artificial intelligence for the real world. https://www.bizjournals.com/boston/news/2018/01/09/hbr-artificial-intelligence-for-the-real-world.html

[26] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (1989), 319–340. https://doi.org/10.2307/249008 Publisher: Management Information Systems Research Center, University of Minnesota.

[27] Aaron M Dollar. 2014. Classifying human hand use and the activities of daily living. *The human hand as an inspiration for robot hand development* (2014), 201–216.

[28] Jason K. Eshraghian. 2020. Human ownership of artificial creativity. 2, 3 (2020), 157–160. https://doi.org/10.1038/s42256-020-0161-x Number: 3 Publisher: Nature Publishing Group.

[29] Loris P Fagioli. 2008. Effects of Illustrations on Retention and Visual Attention Using Authentic Textbooks.

[30] Ian C Fiebelkorn and Sabine Kastner. 2019. A rhythmic theory of attention. *Trends in cognitive sciences* 23, 2 (2019), 87–101.

[31] Paul M Fitts and Michael I Posner. 1967. Human performance. (1967).

[32] Neil D Fleming and Colleen Mills. 1992. Not another inventory, rather a catalyst for reflection. *To improve the academy* 11, 1 (1992), 137–155.

[33] Christopher Frauenberger. 2020. Entanglement HCI The Next Wave? *ACM Transactions on Computer-Human Interaction* 27, 1 (Feb. 2020), 1–27. https://doi.org/10.1145/3364998

[34] Aaron Friedland, Michelle Gilman, Michael Johnson, and Abera Demeke. 2017. Does Reading-While-Listening Enhance Students' Reading Fluency? Preliminary Results from School Experiments in Rural Uganda. *Journal of Education and Practice* 8, 7 (2017), 82–95.

[35] Runshan Fu, Yan Huang, and Param Vir Singh. 2020. AI and Algorithmic Bias: Source, Detection, Mitigation and Implications. https://doi.org/10.2139/ssrn.3681517

[36] Joseph L. Gabbard, J. Edward Swan, II, and Deborah Hix. 2006. The Effects of Text Drawing Styles, Background Textures, and Natural Lighting on Text Legibility in Outdoor Augmented Reality. 15, 1 (2006), 16–32. https://doi.org/10.1162/pres.2006.15.1.16

[37] Linda B. Gambrell and Paula Brooks Jawitz. 1993. Mental Imagery, Text Illustrations, and Children's Story Comprehension and Recall. 28, 3 (1993), 265–276. https://doi.org/10.2307/747998 Publisher: [Wiley, International Reading Association].

[38] Debjyoti Ghosh, Pin Sym Foong, Shengdong Zhao, Can Liu, Nuwan Janaka, and Vinitha Erusu. 2020. EYEditor: Towards On-the-Go Heads-Up Text Editing Using Voice and Manual Input. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376173

[39] Katie L. Glonek and Paul E. King. 2014. Listening to Narratives: An Experimental Examination of Storytelling in the Classroom. 28, 1 (2014), 32–46. https://doi.org/10.1080/10904018.2014.861302 Publisher: Routledge _eprint: https://doi.org/10.1080/10904018.2014.861302.

[40] Xuan Gong, Shuyan Chen, Baochang Zhang, and David Doermann. 2021. Style Consistent Image Generation for Nuclei Instance Segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 3994–4003.

[41] Arthur C. Graesser, Kathy Hauft-Smith, Andrew D. Cohen, and Leonard D. Pyles. 1980. Advanced Outlines, Familiarity, and Text Genre on Retention of Prose. 48, 4 (1980), 281–290. https://doi.org/10.1080/00220973.1980.11011745 Publisher: Routledge.

[42] Melanie C Green and Timothy C Brock. 2000. The role of transportation in the persuasiveness of public narratives. *Journal of personality and social psychology* 79, 5 (2000), 701.

[43] Yike Guo, Qifeng Liu, Jie Chen, Wei Xue, Henrik Jensen, Fernando Rosas, Jeffrey Shaw, Xing Wu, Jiji Zhang, and Jianliang Xu. 2022. Pathway to Future Symbiotic Creativity. *arXiv preprint arXiv:2209.02388* (2022).

[44] Aleesha Hamid, Rabiah Arshad, and Suleman Shahid. 2022. What Are You Thinking?: Using CBT and Storytelling to Improve Mental Health Among College Students. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 441, 16 pages. https://doi.org/10.1145/3491102.3517603

[45] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[46] Iben Have and Birgitte Stougaard Pedersen. 2015. *Digital audiobooks: New media, users, and experiences.* Routledge.

[47] Iben Have and Birgitte Stougaard Pedersen. 2021. *Reading Audiobooks.* Springer International Publishing, Cham, 197–216. https://doi.org/10.1007/978-3-030-49679-1_6

[48] Jeremy M. Heiner, Scott E. Hudson, and Kenichiro Tanaka. 1999. The Information Percolator: Ambient Information Display in a Decorative Object. In *Proceedings of the 12th Annual ACM Symposium on User Interface Software and Technology* (Asheville, North Carolina, USA) *(UIST '99)*. Association for Computing Machinery, New York, NY, USA, 141–148. https://doi.org/10.1145/320719.322595

[49] Sophie K Herbst and Ayelet N Landau. 2016. Rhythms for cognition: the case of temporal processing. *Current Opinion in Behavioral Sciences* 8 (2016), 85–93.

[50] Rebecca J. Hess, Jennifer S. Brach, Sara R. Piva, and Jessie M. VanSwearingen. 2010. Walking Skill Can Be Assessed in Older Adults: Validity of the Figure-of-8 Walk Test. *Physical Therapy* 90, 1 (01 2010), 89–99. https://doi.org/10.2522/ptj.20080121

[51] Kevin Huang, Jiannan Li, Mauricio Sousa, and Tovi Grossman. 2022. Immersive-POV: Filming How-To Videos with a Head-Mounted 360° Action Camera. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 91, 13 pages. https://doi.org/10.1145/3491102.3517468

[52] Kevin Huang, Thad Starner, Ellen Do, Gil Weinberg, Daniel Kohlsdorf, Claas Ahlrichs, and Ruediger Leibrandt. 2010. Mobile music touch: mobile tactile stimulation for passive learning. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 791–800.

[53] Mei lin Jiang and Alex W.C. Tse. 2022. The Effects of Multimedia Glosses, Learning Duration and Working Memory Capacity on High School Students' Passive English Vocabulary Acquisition.

[54] Sebastian Kapp, Michael Barz, Sergey Mukhametov, Daniel Sonntag, and Jochen Kuhn. 2021. ARETT: Augmented Reality Eye Tracking Toolkit for Head Mounted Displays. 21, 6 (2021), 2234. https://doi.org/10.3390/s21062234 Number: 6 Publisher: Multidisciplinary Digital Publishing Institute.

[55] Krzysztof Krejtz, Andrew T. Duchowski, Anna Niedzielska, Cezary Biele, and Izabela Krejtz. 2018. Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. 13, 9 (2018). https://doi.org/10.1371/journal.pone.0203629 Place: US Publisher: Public Library of Science.

[56] Masaaki Kurosu and Kaori Kashimura. 1995. Apparent Usability vs. Inherent Usability: Experimental Analysis on the Determinants of the Apparent Usability. In *Conference Companion on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. Association for Computing Machinery, New York, NY, USA, 292–293. https://doi.org/10.1145/223355.223680

[57] Marc Lambooij, Marten Fortuin, Ingrid Heynderickx, Wijnand IJsselsteijn, Marten Fortuin, Ingrid Heynderickx, and Wijnand IJsselsteijn. 2009. Visual Discomfort and Visual Fatigue of Stereoscopic Displays: A Review. 53 (2009), 1–14. https://doi.org/10.2352/J.ImagingSci.Technol.2009.53.3.030201 Publisher: Society for Imaging Science and Technology.

[58] Lotta C Larson. 2015. E-books and audiobooks: Extending the digital reading experience. *The Reading Teacher* 69, 2 (2015), 169–177.

[59] Kyung-Ryong Lee, Beom Kim, Junyoung Kim, Hwajung Hong, and Young-Woo Park. 2021. ADIO: An interactive artifact physically representing the intangible digital audiobook listening experience in everyday living spaces. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–12.

[60] Joseph Lemley, Anuradha Kar, Alexandru Drimbarean, and Peter Corcoran. 2023. Efficient CNN Implementation for Eye-Gaze Estimation on Low-Power/Low-Quality Consumer Imaging Systems. arXiv:1806.10890 [cs] http://arxiv.org/abs/1806.10890

[61] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. 2019. StoryGAN: A Sequential Conditional GAN for Story Visualization. arXiv:1812.02784 [cs.CV]

[62] Mike E.U. Ligthart, Mark A. Neerincx, and Koen V. Hindriks. 2020. Design Patterns for an Interactive Storytelling Robot to Support Children's Engagement and Agency. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (Cambridge, United Kingdom) *(HRI '20)*. Association for Computing Machinery, New York, NY, USA, 409–418. https://doi.org/10.1145/3319502.3374826

[63] Cindy Lin and Silvia Lindtner. 2021. Techniques of Use: Confronting Value Systems of Productivity, Progress, and Usefulness in Computing and Design. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. https://doi.org/10.1145/3411764.3445237

[64] D. W. Livingstone. 1999. Exploring the Icebergs of Adult Learning: Findings of the First Canadian Survey of Informal Learning Practices. 13, 2 (1999), 49–72. https://cjsae.library.dal.ca/index.php/cjsae/article/view/2000 Number: 2.

[65] Gordon D Logan. 2002. An instance theory of attention and memory. *Psychological review* 109, 2 (2002), 376.

[66] Feiyu Lu, Shakiba Davari, Lee Lisle, Yuan Li, and Doug Bowman. 2020. Glanceable AR: Evaluating Information Access Methods for Head-Worn Augmented Reality. 930–939. https://doi.org/10.1109/VR46266.2020.00113

[67] Victoria J. Marsick and Karen E. Watkins. [n. d.]. Informal and Incidental Learning. 2001, 89 ([n. d.]), 25–34. https://doi.org/10.1002/ace.5 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ace.5.

[68] Tara Matthews. 2006. Designing and Evaluating Glanceable Peripheral Displays. In *Proceedings of the 6th Conference on Designing Interactive Systems* (University Park, PA, USA) *(DIS '06)*. Association for Computing Machinery, New York, NY, USA, 343–345. https://doi.org/10.1145/1142405.1142457

[69] Tara Matthews, Anind K. Dey, Jennifer Mankoff, Scott Carter, and Tye Rattenbury. 2004. A Toolkit for Managing User Attention in Peripheral Displays. In *Proceedings of the 17th Annual ACM Symposium on User Interface Software and Technology* (Santa Fe, NM, USA) *(UIST '04)*. Association for Computing Machinery, New York, NY, USA, 247–256. https://doi.org/10.1145/1029632.1029676

[70] Shannon Maughan. 2023. The audiobook market, and its revenue, Keep Growing. https://www.publishersweekly.com/pw/by-topic/industry-news/audio-books/article/92444-the-audiobook-market-and-revenue-keeps-growing.html

[71] Richard E Mayer. 2002. Multimedia learning. In *Psychology of learning and motivation*. Vol. 41. Elsevier, 85–139.

[72] Microsoft Bing. [n. d.]. *Microsoft Bing Image Creator (Powered by DALL-E)*. https://www.bing.com/create

[73] Natalia Mikidenko and Svetlana Storozheva. 2021. Audiobooks: Reading Practices and Educational Technologies. 97 (2021), 01016. https://doi.org/10.1051/shsconf/20219701016

[74] Tram Thi Minh Tran, Shane Brown, Oliver Weidlich, Mark Billinghurst, and Callum Parker. 2023. Wearable Augmented Reality: Research Trends and Future Directions from Three Major Venues. 29, 11 (2023), 4782–4793. https://doi.org/10.1109/TVCG.2023.3320231

[75] Anca Morar, Maria-Anca Băluțoiu, Alin Moldoveanu, Florica Moldoveanu, and Alex Butean. 2021. CultReal—A Rapid Development Platform for AR Cultural Spaces, with Fused Localization. 21, 19 (2021). https://doi.org/10.3390/s21196618

[76] Roxana Moreno. 2006. Learning in High-Tech and Multimedia Environments. *Current Directions in Psychological Science* 15, 2 (2006), 63–67. https://doi.org/10.1111/j.0963-7214.2006.00408.x

[77] Elizabeth L. Murnane, Yekaterina S. Glazko, Jean Costa, Raymond Yao, Grace Zhao, Paula M. L. Moya, and James A. Landay. 2023. Narrative-Based Visual Feedback to Encourage Sustained Physical Activity: A Field Trial of the WhoIsZuki Mobile Health Platform. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 1, Article 23 (mar 2023), 36 pages. https://doi.org/10.1145/3580786

[78] Janet H Murray. 2017. *Hamlet on the Holodeck, updated edition: The Future of Narrative in Cyberspace.* MIT press.

[79] Ali Neshati, Yumiko Sakamoto, Launa Leboe-Mcgowan, Jason Leboe-Mcgowan, Marcos Serrano, and Pourang Irani. 2019. G-Sparks: Glanceable Sparklines on Smartwatches. In *45th Conference on Graphics Interface (GI 2019)*. Kingston, Ontario, Canada, 1–9. https://doi.org/10.20380/GI2019.23

[80] Rachel Noorda and Kathi Inman Berens. 2021. Immersive Media and Books 2020: New Insights About Book Pirates, Libraries and Discovery, Millennials, and Cross-Media Engagement: Before and During COVID.

[81] Donald A Norman. 2004. *Emotional design: Why we love (or hate) everyday things.* Civitas Books.

[82] The Editors of Goop. 2015. Konmari method - Marie Kondo folding guide for clothes | Goop. https://goop.com/food/decorating-design/the-illustrated-guide-to-the-kondo-mari-method/

[83] Eyal Ophir, Clifford Nass, and Anthony D. Wagner. 2009. Cognitive control in media multitaskers. 106, 37 (2009), 15583–15587. https://doi.org/10.1073/pnas.0903620106 _eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.0903620106.

[84] Josetxu Orrantia, David Múñez, and Julio Tarín. 2014. Connecting goals and actions during reading: The role of illustrations. *Reading and Writing* 27 (2014), 153–170.

[85] Anna-Marie Ortloff, Lydia Güntner, Maximiliane Windl, Thomas Schmidt, Martin Kocur, and Christian Wolff. 2019. SentiBooks: Enhancing Audiobooks via Affective Computing and Smart Light Bulbs. In *Proceedings of Mensch Und Computer 2019* (Hamburg, Germany) *(MuC '19)*. Association for Computing Machinery, New York, NY, USA, 863–866. https://doi.org/10.1145/3340764.3345368

[86] Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie* 45, 3 (1991), 255.

[87] Renate Prins, Lucy Avraamidou, and Martin Goedhart. 2017. Tell me a Story: the use of narrative as a learning tool for natural selection. 54, 1 (2017), 20–33. https://doi.org/10.1080/09523987.2017.1324361 Publisher: Routledge _eprint: https://doi.org/10.1080/09523987.2017.1324361.

[88] Ashwin Ram and Shengdong Zhao. 2021. LSVP: Towards Effective On-the-Go Video Learning Using Optical Head-Mounted Displays. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 30 (mar 2021), 27 pages. https://doi.org/10.1145/3448118

[89] Ashwin Ram and Shengdong Zhao. 2022. Does Dynamically Drawn Text Improve Learning? Investigating the Effect of Text Presentation Styles in Video Learning. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 89, 12 pages. https://doi.org/10.1145/3491102.3517499
[90] Luz Rello and Ricardo Baeza-Yates. 2013. Good fonts for dyslexia. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility* (Bellevue Washington, 2013-10-21). ACM, 1–8. https://doi.org/10.1145/2513383.2513447
[91] James A. Rose. [n. d.]. To Teach Science, Tell Stories. https://www.semanticscholar.org/paper/To-Teach-Science%2C-Tell-Stories-Rose/8dcd13bbd719b1be288c8b3c60542f74c41411b6
[92] Louise M. Rosenblatt. [n. d.]. The Transactional Theory: Against Dualisms. 55, 4 ([n. d.]), 377. https://doi.org/10.2307/378648
[93] David Rothman. 2020. Most ebook and Audiobook Users Multitask: New study explores the details. https://teleread.org/2020/11/06/most-ebook-and-audiobook-users-multitask-new-study-explores-the-details/
[94] D.C. Rubin. 1995. *Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes*. Oxford University Press. https://books.google.com/books?id=yq5A7djClegC
[95] Rufat Rzayev, Paweł W. Woźniak, Tilman Dingler, and Niels Henze. 2018. Reading on Smart Glasses: The Effect of Text Position, Presentation Type and Walking. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3173574.3173619
[96] Dario D. Salvucci. 2013. Multitasking. In *The Oxford handbook of cognitive engineering*. Oxford University Press, 57–67. https://doi.org/10.1093/oxfordhb/9780199757183.001.0001
[97] Pamela Samuelson. 2023. Generative AI meets copyright. 381, 6654 (2023), 158–161. https://doi.org/10.1126/science.adi0656 Publisher: American Association for the Advancement of Science.
[98] Shardul Sapkota, Ashwin Ram, and Shengdong Zhao. 2021. Ubiquitous Interactions for Heads-Up Computing: Understanding Users' Preferences for Subtle Interaction Techniques in Everyday Settings. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction* (Toulouse & Virtual, France) *(MobileHCI '21)*. Association for Computing Machinery, New York, NY, USA, Article 36, 15 pages. https://doi.org/10.1145/3447526.3472035
[99] Diane Lemonnier Schallert. 1980. *The role of illustrations in reading comprehension*. Lawrence Erlbaum, 503–525.
[100] Diane Lemonnier Schallert. 2017. The role of illustrations in reading comprehension. *Theoretical issues in reading comprehension* (2017), 503–524.
[101] Daniel Schugurensky. 2000. The Forms of Informal Learning: Towards a Conceptualization of the Field. (2000).
[102] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.
[103] Daniel Schwajda, Judith Friedl, Fabian Pointecker, Hans-Christian Jetter, and Christoph Anthes. 2023. Transforming graph data visualisations from 2D displays into augmented reality 3D space: A quantitative study. 4 (2023). https://www.frontiersin.org/articles/10.3389/frvir.2023.1155628
[104] Jae-eun Shin, Boram Yoon, Dooyoung Kim, and Woontack Woo. 2021. A User-Oriented Approach to Space-Adaptive Augmentation: The Effects of Spatial Affordance on Narrative Experience in an Augmented Reality Detective Game. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 722, 13 pages. https://doi.org/10.1145/3411764.3445675
[105] Anisha Singh and Patricia A Alexander. 2022. Audiobooks, print, and comprehension: What we know and what we need to know. *Educational Psychology Review* 34, 2 (2022), 677–715.
[106] Jacob Somervell, D. Scott McCrickard, Chris North, and Maulik Shukla. 2002. An Evaluation of Information Visualization in Attention-Limited Environments. In *Eurographics / IEEE VGTC Symposium on Visualization*, D. Ebert, P. Brunet, and I. Navazo (Eds.). The Eurographics Association. https://doi.org/10.2312/VisSym/VisSym02/211-216
[107] Ramya Srinivasan and Devi Parikh. 2021. Building Bridges: Generative Artworks to Explore AI Ethics. *arXiv preprint arXiv:2106.13901* (2021).
[108] Ramya Srinivasan and Kanji Uchino. 2021. Biases in Generative Art: A Causal Look from the Lens of Art History. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2021-03-01) *(FAccT '21)*. Association for Computing Machinery, 41–51. https://doi.org/10.1145/3442188.3445869
[109] Felicia Fang-Yi Tan, Ashwin Ram, Chloe Haigh, and Shengdong Zhao. 2023. Mindful Moments: Exploring On-the-go Mindfulness Practice On Smart-glasses. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference* (<conf-loc>, <city>Pittsburgh</city>, <state>PA</state>, <country>USA</country>, </conf-loc>) *(DIS '23)*. Association for Computing Machinery, New York, NY, USA, 476–492. https://doi.org/10.1145/3563657.3596030
[110] Xuli Tang, Xin Li, Ying Ding, Min Song, and Yi Bu. 2020. The pace of artificial intelligence innovations: Speed, talent, and trial-and-error. *Journal of Informetrics* 14, 4 (2020), 101094. https://doi.org/10.1016/j.joi.2020.101094
[111] Laura Tarantino, Daniela Angelucci, Alessandra Bonomo, Annalisa Cardinali, and Stefania Di Paolo. 2021. Design and applications of GLANCE: GLanceable alarm notification for a user centered experience. *Applied Sciences* 11, 2 (2021), 669.
[112] Elisa Tattersall Wallin. 2022. Audiobook apps: exploring reading practices and technical affordances in the player features. *Information research* 27, 4 (2022).
[113] Edward R Tufte. 1991. Envisioning information. *Optometry and Vision Science* 68, 4 (1991), 322–324.
[114] Jacob Turner. 2019. Responsibility for AI. In *Robot Rules : Regulating Artificial Intelligence*, Jacob Turner (Ed.). Springer International Publishing, 81–132. https://doi.org/10.1007/978-3-319-96235-1_3
[115] Unity Real-Time Development Platform. [n. d.]. *Unity*. https://unity.com/
[116] Karl Werder, Balasubramaniam Ramesh, and Rongen (Sophia) Zhang. 2022. Establishing Data Provenance for Responsible Artificial Intelligence Systems. 13, 2 (2022), 22:1–22:23. https://doi.org/10.1145/3503488
[117] Christopher D Wickens. 2002. Multiple resources and performance prediction. *Theoretical issues in ergonomics science* 3, 2 (2002), 159–177.
[118] Christopher D. Wickens, Michael Vidulich, and Diane Sandry-Garza. 1984. Principles of S-C-R Compatibility with Spatial and Verbal Tasks: The Role of Display-Control Location and Voice-Interactive Display-Control Interfacing. *Human Factors* 26, 5 (1984), 533–543. https://doi.org/10.1177/001872088402600505
[119] Michael B. W. Wolfe and Joseph A. Mienko. 2007. Learning and memory of factual content from narrative and expository text. 77, 3 (2007), 541–564. https://doi.org/10.1348/000709906X143902 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1348/000709906X143902.
[120] Bang Wong. [n. d.]. Points of view: Color blindness. 8, 6 ([n. d.]), 441–441. https://doi.org/10.1038/nmeth.1618
[121] Marjorie Woollacott and Anne Shumway-Cook. 2002. Attention and the control of posture and gait: a review of an emerging area of research. *Gait & Posture* 16, 1 (2002), 1–14. https://doi.org/10.1016/S0966-6362(01)00156-4
[122] Xreal. [n. d.]. https://www.xreal.com/light/
[123] Dominyk Zdanovic, Tanja Julie Lembcke, and Toine Bogers. [n. d.]. The Influence of Data Storytelling on the Ability to Recall Information. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (New York, NY, USA, 2022-03-14) *(CHIIR '22)*. Association for Computing Machinery, 67–77. https://doi.org/10.1145/3498366.3505755
[124] Fang Zhao, Robert Gaschler, Anneli Kneschke, Simon Radler, Melanie Gausmann, Christina Duttine, and Hilde Haider. 2020. Origami folding: Taxing resources necessary for the acquisition of sequential skills. *Plos one* 15, 10 (2020), e0240226.
[125] Shengdong Zhao, Felicia Tan, and Katherine Fennedy. 2023. Heads-Up Computing Moving Beyond the Device-Centered Paradigm. *Commun. ACM* 66, 9 (Aug 2023), 56–63. https://doi.org/10.1145/3571722

A LLM-ASSISTED WORKFLOW FOR NARRATIVE ILLUSTRATION GENERATION

We offer a step-by-step approach that researchers and designers can follow to generate AI-assisted visual augmentations for audiobooks. We believe it provides a starting point for those interested in building on AudioXtend, noting however, that this should be updated in accordance to the rapidly evolving landscape of AI, with new research articles, methodologies and tools frequently released [110].

Step 1: Text Segmentation

Start by presenting the LLM (we used GPT-4) with the audiobook text segment that you plan to illustrate.

> Prompt: The following text is exacted from a book. Please read through it and I will ask questions based on it: «Insert text»

Ask the LLM to divide the text into parts, each deserving of its own illustration. Here, you can specify the number of parts you wish to have, as determined by the total illustrations you wish to have in the set.

> Prompt: Separate the original paragraph into «X no. of total illustrations» parts based on the progression of the story. Each part should have a title with keywords.

The task of partitioning the text into segments that effectively capture its narrative essence is both subjective and complex, varying from one designer to another. We offer some theoretical and practical resources which the reader can refer to for inspiration and guidance, such as narratological principles as outlined by Bal [6], Carney and Levin's "Guidelines for Educators Considering Text-Accompanying Illustrations," [19] and criteria for identifying text-relevant illustrations as developed by Schallert [100].

Step 2: Art Style Recommendations

To decide on an art style that matches the tone and content of the audiobook, inquire about suitable art styles for the story segment. These provide options for designs which you can test.

> Prompt: Recommend and describe three art styles for this particular story segment, from the most informative to the most minimalistic one. Excluded styles: «excluded styles»

As understood from earlier discussions, consider grayscale styles to minimize visual obstruction. For OHMD hardware such as the Xreal Light, black color on screen appears transparent, thus, a grayscale illustration is more likely to provide simpler illustrations.

> Prompt: Only in black and white, grayscale, what are two styles that are very different?

Step 3: Content Summarization

To capture multi-dimensional features of each text segment, including photography angles, sentiment, interactions between characters, etc.

> Prompt: Summarize each part in terms of: 1. Two adjectives to describe the sentiment. 2. Type and angle of shot in photography terms. 3. The title of the part. 4. One sentence that describes the background environment. 5. One sentence that describes the key characters and their interaction, if any. 6. Highlight the key objects, if any. Finally, create a table.

At this point, check the table for alignment with your creative objectives.

Step 4: Image Generation Prompts

Create image prompts that can be fed into an AI image-generation model. Specify the art style of your choice from Step 2.

> Prompt: Now you are communicating with an image generation AI. Construct prompts for «X no. of total illustrations» illustrations based on these parts, in the style of «art style». Note that each prompt must be self-contained as the AI processes them individually. Remove instructions like "Generate an image".

Step 5: Prompt Simplification

Simplify the generated prompts to their essence for efficient image generation.

> Prompt: Simplify each prompt to include only the most relevant information and keywords.

Repeat this prompt for further simplification. After arriving at an optimized prompt, input it into your chosen AI image-generator tool. Note that the initial output may necessitate fine-tuning, either by using AI tool options or through manual editing on a graphics software. Pilot testing mid-way through an illustration set with a small audience (as was done in all parts of our research) can provide insights into the effectiveness of the AI-generated illustrations before a fuller rollout. This iterative refinement ensures that the final illustrations closely align with the narrative requirements.