ImReader: A Multimodal Tool for Immersive Narrative Reading Utilizing LLMs and Generative Models

Sitong Pan
The college of Earth Science, Zhejiang University
Hangzhou, China
pstong@zju.edu.cn

Biying Xu
School of Software Technology, Zhejiang University
Hangzhou, China
2094197975@qq.com

Fangyuan Ye
The School of International Studies, Zhejiang University
Hangzhou, China
3210102071@zju.edu.cn

ABSTRACT
With technological advancements and changing digital content consumption habits, more readers are now seeking richer and more diverse reading experiences. Multimodal content can significantly enhance the reading experience of static texts. While large language models (LLMs) and generative models open up new opportunities for generating multimodal content, no research has yet applied them to reading technology. Therefore, this article introduces ImReader, a multimodal reading tool designed for immersive storytelling based on LLMs and generative models. Specifically, ImReader generates illustrations, sound effects, background music, and other audiovisual content that matches the user-specified reading text. Users can also customize their preferences using tags. This is achieved by offering pre-defined style options and user preferences input. In the preliminary test, ImReader has demonstrated improved and enriched reading experiences.

Author Keywords
Multi-modal generation; Large Language Model(LLM); Generative Model; Reading Interest; E-book

1 INTRODUCTION
In today's digital age, a growing number of individuals worldwide are embracing mobile devices and e-readers as their preferred means of reading.[1]Despite advances in digital reading technologies, most current applications remain primarily focused on text display and basic interactivity, such as highlighting and note-taking. These applications rarely leverage the full potential of today's diverse and efficient generative models to create a more immersive and engaging reading experience. However, there is an increasing demand for more diverse elements, such as generated images and sounds, in digital reading platforms to enhance the overall reading experience beyond basic functionalities. Experiencing multisensory stimulation is a method of eliciting immersion. [2]

UbiComp '13, Sept 9-12, 2013, Zurich, Switzerland.

Recently,the emergence of generative models large language models have opened new avenues for immersive audiovisual experiences which could be applied in reading technology because of the following reasons:

1. Generative ability: Generative models, known for their ability to create diverse and rich content, hold significant potential in generating images and audios that correspond to input prompt. This makes transforming digital reading by integrating multimedia elements possible. [3]

2. Semantic ability: The quality and relevance of the content generated by generative models are heavily influenced by the prompts they receive. This is where large language models (LLMs) come into play. LLMs possess powerful semantic generation capabilities, enabling them to craft nuanced and contextually appropriate prompts for generative models. Thus, our project aims to harness the semantic prowess of LLMs in conjunction with generative models to elevate digital reading platforms beyond their current limitations, aiming to enrich users' reading experiences. [4]

This article introduces ImReader, a novel multimodal reading application that leverages large language models' and generative models' capabilities to provide a more immersive and personalized reading experience that goes beyond static text. Specifically, by integrating GPT-4 and Suno, this reading application generates images and music on the content underlined by readers, thereby enriching both cognitive and emotional aspects of reading. ImReader offers users the ability to customize the style of generated images (e.g., cartoon, realistic) and music (e.g., gentle, pop), transforming passive reading into an interactive and engaging activity.

2 RELATED WORK
The significance of enhancing user immersion and user experience in multimedia applications is widely acknowledged and has been applied across various fields, including education and entertainment.[5][6]Prior research has explored various approaches to enrich the reading experience through multi-sensory stimulation.[7] Alam et al. (2013) propose the HE-Book system, which integrates hapto-audio-visual interactions with home entertainment systems and haptic devices to achieve an immersive reading space. [8] Similarly, Sanchez et al. (2016) developed an augmented e-Reader prototype that creates an embodied reading experience using

sound and haptic feedback. [9]Ellen et al. (2021) introduce MBook as a tool to present multi-sensorial books, impacting the quality of experience and showing good usability among students.[10] However, in previous multimodal readers, multimodal content creation required manual coding and annotation, which resulted in high costs, insufficient content personalization, and too strong restrictions. This article introduced the generative model to solve this problem.

The semantic capabilities of Large Language Models (LLMs) have been utilized in various domains for multimodal content creation. Imreader leverages the semantic capabilities to produce prompts for such content. Similarly, numerous studies have employed LLMs to generate descriptive texts for multimodal content creation. Shen et al. (2024) used LLMs to extract user preference descriptions from user behaviors, enabling personalized multimodal generation.[11] He et al. (2023) utilized an LLM engine to interpret inputs, translating abstract notions into specific instructions for subsequent generation modules. [12]Zhao et al. (2023) employed LLMs to generate detailed descriptions and cues from text contexts, facilitating visual responses in diffusion models. [13]Yang et al. (2023) used GPT-4V to generate multiple text prompts corresponding to a user's multimodal input, conditioned on previous text feedback and refinement history.[14] Additionally, Yu et al. (2023) applied LLMs to brainstorm prompts for high-quality image synthesis based on label words and user descriptions. While these works primarily focus on image generation, the application of such methods in music generation remains less common. [15]Our research extends this approach to music generation, demonstrating its potential in this new domain.

3 SYSTEM ARCHITECTURE

In this section we will introduce the system architecture of ImReader, from the aspects of system characteristics and procedure.

3.1 Generative Model and LLM-powered Multi-modal Reader

ImReader is a mobile reader which generates multi-modal output for selected narration contents to provide users with immersive reading experiences. The system is driven by a large language model(GPT-4) and generative modesl(Suno and DALL-E). While reading, users can select any text they wish to extend, and press the "Generate" button. ImReader will then produce corresponding illustrations or music that represent the narrative content visually or audibly, tailored to the user's predefined style preferences. Upon importing a book, the text is parsed into three hierarchical levels: book, chapter, and paragraph. This structure allows for generating content with varying levels of detail, ensuring that the output is comprehensive and contextually appropriate for each level. There are three significant features and considerations in our system:

3.1.1 Context consistence

When users utilize Large Language Models (LLMs) to generate content in specific ways, they often have both "explicit needs" and "implicit needs." The former directly relates to the generated object, whereas the latter supports the former by providing necessary background or stylistic context. For instance, when a user wants to generate an illustration for a specific scene described in the text, it is insufficient to rely solely on the selected text. Instead, we should also consider the entire paragraph or chapter in which the text appears. This approach ensures that the generated illustration aligns with the overall style and atmosphere of the narrative, thus enhancing the coherence and relevance of the visual representation.

3.1.2 User intervention

1. User control over generated content

   In the interview regarding the generation mode of content, specifically asking whether the content should be "automatically" or "manually" generated, a significant majority of interviewees expressed a preference for manual activation. This preference also extended to the methods for presenting multimodal content; for example, the majority of users favored initiating music playback manually instead of automatically upon its creation. Additionally, when discussing the display of generated illustrations, the preferred approach was to discreetly place the content as a small icon at the end of the paragraph, allowing users to access it manually if they choose to do so.

   > "I prefer to manually initiate music playback. It could be quite embarrassing if the reader automatically played audio in a public setting without headphones. Moreover, when I am engrossed in reading, unexpected sounds can startle me easily."-Interviewee 1

   > "I'm not opposed to auto-play, but I would require a clear indication before it begins, even if it slightly interrupts my reading. Additionally, it is crucial to have the capability to pause the music or any other multimodal effects at any moment."-Interviewee 2

   These findings underscore a general preference for less obtrusive content delivery methods which was adopted by ImReader system. To elaborate further, generated illustrations are represented by an icon initially. Only when a user clicks on this icon will the illustration enlarge for detailed viewing. Similarly, the music follows the same principle; it only plays when a user opts to initiate playback, rather than automatically playing upon generation. The design id expected to minimize disruptions to users' reading flow and provide user-controlled interaction within digital reading platforms to enhance user satisfaction and engagement.

2. Customizable preferences

   In our survey, respondents displayed a variety of preferences and expectations, ranging from basic reader appearance settings to multi-modal content style preferences. Therefore, ImReader provide s users customizable preference settings to meet these diverse needs. In addition to global settings, users can manually specify different multi-modal content styles for each book. Specifically, they can select or manullay type their desired style 'Tags' (e.g.,Cartoon, Realistic for illustration, Quiet, Rock for

music), which will be integrated into prompts, therefore affecting the generated image and music . Such design offers users more flexibility to meet with their personalized needs.

3.1.3 Large language model and generative model based

In the implementation of ImReader, distinct approaches were adopted for generating visual and auditory content due to their inherent differences in abstraction and representational styles. For visual content, a static prompt template is used to generate images directly through DALL-E 2, an image generation model by OpenAI. This approach is effective as images need to concretely depict the content of the text, providing a direct visual representation of the narrative elements.

Conversely, music, by its nature, encapsulates more abstract qualities such as mood and atmosphere, which are not as straightforwardly derived from text. To address this, we employ GPT-4 to generate descriptive prompts that articulate the intended musical style and ambiance. These descriptions are generated based on the text selected by the user, coupled with predefined style tags that reflect the user's aesthetic preferences. The resulting text prompts are then used as input for Suno, a music synthesis AI, allowing for the creation of music that complements the reading experience by enhancing the emotional and atmospheric depth of the narrative."

In the development of ImReader, we employed a two-pronged approach to enrich the reading experience through both visual and auditory elements. For the visual component, we utilized a static prompt template to directly generate illustrations using DALL-E 2, an advanced image synthesis model embedded within GPT-4. This method ensures consistency and relevancy in the visual representations of the text. On the auditory side, we leveraged GPT-4 to generate dynamic textual prompts based on the selected text by the user. These prompts were then used as inputs for Suno, a text-to-music AI model, to produce music that audibly reflects the narrative's mood and themes. This dual approach allows ImReader to provide a comprehensive sensory experience by visually and audibly representing the narrative content.

3.2 Procedure

The whole procedure can be divided into two main phases: preference setting phase and reading phase, as seen in Figure 1.

3.2.1 Preference Setting Phase

Set comfortable global reading configurations and personalized multi-modal preferences. In this phase of the application, users are allowed to adjust foundational reading settings, such as font type, background color, line spacing, and margins. Additionally, users can set their preferences for multi-modal content. As illustrated in Figure [], we provide a selection of predefined tags for images and music styles from which users can choose. They also have the option to enter and add their own custom tags into the tag bar located below. These tags are then integrated with the selected text during reading to generate corresponding musical or visual content.

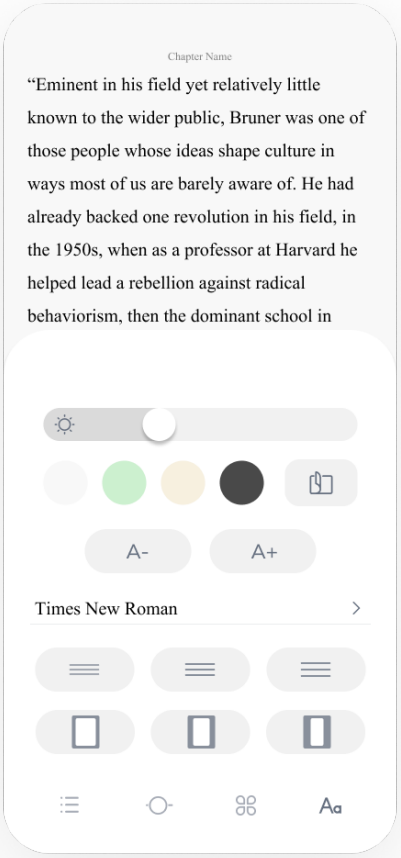

Figure 1: Users can select tags or input tags

This feature enhances the user's experience by allowing personalized adjustments that reflect their individual reading and sensory preferences.

3.2.2 Reading Phase

Multi-modal generation by selecting sentences. In this phase, users can select sentences or entire paragraphs during their reading sessions and choose to generate corresponding images or music. Within the ImReader system, the methods for generating images and music differ. However, both methods share a common foundation: they combine the user's selected text, the context surrounding the selected content, and predefined style tags.

1. **Image Generation** Images are particularly effective at visually representing and illustrating the content described in the text. To leverage this, we use a prompt template that has been refined through experimentation to yield optimal results. This template integrates the selected text and its surrounding context with the predefined style tags to form a coherent prompt. This prompt is then input into the

DALL-E model embedded within GPT-4 to generate the image. The prompt used for image generation is designed to ensure that the generated image accurately reflects the main content of the selected text while maintaining consistency with the broader narrative context. The specific prompt template is as follows: "Please generate an image with the following text as the content, and the image style should match the text style: "$text" +"You can also refer to the context here to infer other elements in the picture, but these should not be the main content of the image: $context;"

This approach ensures that the generated image is both contextually relevant and stylistically aligned with the user's preferences, providing a visually rich enhancement to the reading experience.

2. **Music Generation** Music generation differs significantly from image generation in that it is inherently more abstract and holistic. Specifically, music conveys emotions, atmosphere, and moods rather than representing specific textual scenes or characters. Consequently, generating music requires more detailed and contextually rich textual descriptions to guide the generation model effectively. Our strategy leverages the semantic capabilities of large language models to create text descriptions that align with the content and mood of the selected text, which are then used as prompts for the music generation model. This process involves three detailed steps:

   1. Generating a Musical Style Description: The first step involves inputting the selected text and its context into a large language model to generate a concise description of the musical style. The prompt used for this step is as follows: "Please summarize the musical melody and style of this passage. Please try to describe them in a concise phrase. Do not include anything else in your answer: $text "

   2. Creating a Comprehensive Music Description Prompt: In the second step, we use the large language model to take the initial AI-generated prompt from the first step and combine it with predefined stylistic tags. This results in a detailed and stylistically aligned description of the music, which serves as the input for the music generation model. The prompt for this step is: "The music has the following styles: $input + $tags.join(", "), and please add appropriate lyrics or no lyrics."

   3. Generating the Music: The final step involves combining the prompts generated in the first two steps and inputting them into the Suno music generation model to produce the final musical output.

   This three-step process ensures that the generated music is contextually relevant and aligns with the user's stylistic preferences. By integrating the selected text, its context, and predefined style tags, the approach enhances the overall reading experience by providing a well-matched auditory component. This method demonstrates a nuanced understanding of the interplay between textual content and musical interpretation, aiming to create a more immersive and personalized user experience.

4 PRELIMINARY USE CASE

In this chapter, we present a preliminary use case to demonstrate the functionality and effectiveness of the ImReader system. The system allows users to enhance their reading experience by generating contextually appropriate images and music based on selected text passages. The following sections describe the process and outcomes of using ImReader for a sample text excerpt.

4.1 Use Case Scenario

Consider a user reading a novel. The user comes across a particularly vivid paragraph that describes a serene forest scene. The user decides to generate both an image and music to enhance the reading experience.

Selected Text: "The sun filtered through the dense canopy of leaves, casting dappled shadows on the forest floor. Birds chirped melodiously, and a gentle breeze rustled the foliage, creating a symphony of nature. The air was fresh with the scent of pine and earth, and the tranquility of the forest was palpable."

4.2 Generating the Image

To generate the image, the selected text and its context are input into the ImReader system. The system uses a prompt template to ensure the generated image aligns with the textual description and overall mood.

Image Generation Prompt: "Please generate an image with the following text as the content, and the image style should match the text style: The sun filtered through the dense canopy of leaves, casting dappled shadows on the forest floor. Birds chirped melodiously, and a gentle breeze rustled the foliage, creating a symphony of nature. The air was fresh with the scent of pine and earth, and the tranquility of the forest was palpable. You can also refer to the context here to infer other elements in the picture, but these should not be the main content of the image: Before reaching the forest, the protagonist had been wandering through a dense and bustling city, overwhelmed by the noise and chaos.....";

4.3 Generating the Music

Music generation is handled differently, as it requires capturing the mood and atmosphere conveyed by the text. The process involves creating a detailed textual description of the music style, which is then used to generate the music.

Step 1: Generate Musical Style Description

"Please summarize the musical melody and style of this passage. Please try to describe them in a concise phrase. Do not include anything else in your answer: "The sun filtered through the dense canopy of leaves, casting dappled shadows on the forest floor. Birds chirped melodiously, and a gentle breeze rustled the foliage, creating a symphony of nature. The air was fresh with the scent of pine and earth, and the tranquility of the forest was palpable." Generated Description: "Tranquil and melodious, evoking nature's symphony with gentle, harmonic overtones."

Step 2: Create Detailed Music Description Prompt

Figure 2: The generated illustration

”The music has the following styles: Tranquil and melodious, evoking nature’s symphony with gentle, harmonic overtones, and please add appropriate lyrics or no lyrics.”

Step 3: Generate the Music

Using the combined prompts, the final music is generated using the Suno music generation model.

4.4 Conclusion

This preliminary use case demonstrates the capabilities of the ImReader system in enhancing the reading experience through multi-modal content generation. By generating images and music that align with the text’s context and user preferences, ImReader offers a richer, more immersive experience for readers. This approach underscores the potential of integrating advanced AI models in digital reading applications to meet the diverse needs and preferences of users.

5 CONCLUSION

This article proposes ImReader, a prototype multimodal reading application with generated music and images. By employing GPT-4, DALL-E and Suno, ImReader offers an enriched and personalized reading experience by integrating auditory and visual content. This approach has the potential to transform traditional reading by making it more immersive and engaging, addressing the limitations of current digital reading applications.

The implications and effectiveness of integrating generated contents into multi-modal reading application need to be investigated. Future research should focus on assessing user experience and the cognitive benefits of such technologies. Key areas of interest include memory retention, emotional engagement, and overall satisfaction with the reading experience. Also, integrating multimedia content without disrupting the reading process is challenging. Excessive sensory input can overwhelm users, thereby affecting their reading experience. Thus, the impact of different multimedia integration methods on the reading experience could be explored. Understanding these aspects can contribute to the advancement of reading technology.Moreover, The capabilities of the generative models need to be enhanced to ensure that the generated auditory and visual stimuli are closely aligned with the narrative and emotional tone of the text.

In addition to auditory and visual modalities, future expansions could include other sensory modalities and environmental interactions. For instance, haptic feedback could provide tactile sensations corresponding to story events and emotional tones. Environmental interaction could be achieved by utilizing the device’s sensors and augmented reality (AR). Furthermore, wearable devices could monitor the user’s physiological responses and adjust the content accordingly, creating a more tailored and immersive experience.

Finally, by examining the impact of generative models on the reading experience, we aim to provide insights into the potential of multi-modal reading technologies for educational and recreational purposes and create a better reading experience for users across various contexts.

REFERENCES

1. Somipam Shimray, Chennupati Keerti, and Chennupati Ramaiah. An overview of mobile reading habits. *DESIDOC Journal of Library Information Technology*, 35:364–375, 10 2015.
2. agrawal sarvesh, simon adèle, bech søren, bæntsen klaus, and forchhammer søren. defining immersion: literature review and implications for research on audiovisual experiences. *journal of the audio engineering society*, 68:404–417, june 2020.
3. Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. Multimodal large language models: A survey, 2023.
4. Yilun Zhu, Joel Ruben Antony Moniz, Shruti Bhargava, Jiarui Lu, Dhivya Piraviperumal, Site Li, Yuan Zhang, Hong Yu, and Bo-Hsiang Tseng. Can large language models understand context?, 2024.
5. Péter Tamás Kovács, Niall Murray, Gregor Rozinaj, Yevgeniya Sulema, and Renata Rybárová. Application of immersive technologies for education: State of the art. In *2015 International Conference on Interactive Mobile Communication Technologies and Learning (IMCL)*, pages 283–288, 2015.
6. Stefano Cacciaguerra, Marco Roccetti, and Paola Salomoni. *Multimedia Entertainment Applications*, pages 510–518. Springer US, Boston, MA, 2006.
7. Pedro Ribeiro, Wolfgang Müller, Ido Iurgel, Christian Ressel, and Carrie Ching. A design space for digital augmentation of reading. In Constantine Stephanidis, Margherita Antona, Stavroula Ntoa, and Gavriel Salvendy, editors, *HCI International 2023 – Late*

Breaking Posters, pages 200–208. Springer Nature Switzerland, 2024.

8. Kazi Masudul Alam, Abu Saleh Md Mahfujur Rahman, and Abdulmotaleb El Saddik. Mobile haptic e-book system to support 3d immersive reading in ubiquitous environments. *ACM Trans. Multimedia Comput. Commun. Appl.*, 9(4), aug 2013.
9. Susana Sanchez, Tilman Dingler, Heng Gu, and Kai Kunze. Embodied reading: A multisensory experience. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, page 1459–1466, New York, NY, USA, 2016. Association for Computing Machinery.
10. Ellen P. Silva, Natália Vieira, Glauco Amorim, Renata Mousinho, Gustavo Guedes, Gheorghita Ghinea, and Joel A. F. Dos Santos. Using multisensory content to impact the quality of experience of reading digital books. *ACM Trans. Multimedia Comput. Commun. Appl.*, 17(4), nov 2021.
11. Xiaoteng Shen, Rui Zhang, Xiaoyan Zhao, Jieming Zhu, and Xi Xiao. Pmg : Personalized multimodal generation with large language models, 2024.
12. Jun-Yan He, Zhi-Qi Cheng, Chenyang Li, Jingdong Sun, Wangmeng Xiang, Xianhui Lin, Xiaoyang Kang, Zengke Jin, Yusen Hu, Bin Luo, Yifeng Geng, Xuansong Xie, and Jingren Zhou. Wordart designer: User-driven artistic typography synthesis using large language models, 2023.
13. Xiangyu Zhao, Bo Liu, Qijiong Liu, Guangyuan Shi, and Xiao-Ming Wu. Easygen: Easing multimodal generation with bidiffuser and llms, 2024.
14. Zhengyuan Yang, Jianfeng Wang, Linjie Li, Kevin Lin, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. Idea2img: Iterative self-refinement with gpt-4v(ision) for automatic image design and generation, 2023.
15. Qifan Yu, Juncheng Li, Wentao Ye, Siliang Tang, and Yueting Zhuang. Interactive data synthesis for systematic vision adaptation via llms-aigcs collaboration, 2023.