Special Issue on Consumer Electronics

Gen-Presenter: An Interactive AIoT Presentation System with RAG-Supported LLMs

Ming-Shun Wang, Showkat Ahmad Bhat, Jian-Feng Li, Ming-Che Chen
Southern Taiwan University of Science and Technology

***Abstract*—Gen-Presenter is a cost-effective, AI-powered interactive presentation system designed for consumer-facing environments such as museums, exhibitions, and public installations. Addressing the limitations of traditional kiosks and robotic guides, it leverages Retrieval-Augmented Generation (RAG), large language models (LLMs), and edge AI to deliver adaptive multimodal content tailored to visitor demographics and engagement levels. Running primarily on an edge device, the system integrates modules for gaze detection, age estimation, automated speech recognition (ASR), text-to-speech (TTS), and content management. Two LLMs on a backend NLP server handle slide selection and explanation generation. While LLM2 with RAG enriched responses with contextual content, LLM1 (LLaMA 3.3 without RAG) demonstrated superior slide selection performance, achieving 0.95, 0.91, and 0.96 accuracy with only 1s of latency. User suitability testing demonstrates strong alignment with diverse age groups (scores between 0.81 to 0.96), and the multithreaded design delivers a 15.7x improvement in slide processing efficiency. These results highlight the system's scalability, responsiveness, and potential as an intelligent consumer electronics solution for immersive public engagement.**

IMAGINE WALKING through an exhibit where a smart display engages each visitor individually, answering questions in real time and adapting to their interests. With advances in AI and IoT, such responsive systems are becoming increasingly feasible, offering a dynamic alternative to conventional kiosks that deliver static, one-size-fits-all content [1]–[2]. In public spaces like museums, where audiences span diverse age groups, cultural backgrounds, and engagement levels, static displays often fall short to provide meaningful interaction [3]. Recent studies report that 74% of museum visitors find personalized digital signage sig-

Digital Object Identifier 10.1109/MCE.2022.Doi Number
Date of publication 00 xxxx 0000; date of current version 00 xxxx 0000

nificantly more engaging than generic alternatives [4], highlighting a growing demand for tailored, interactive experiences. In response, researchers and developers are exploring consumer-facing systems that leverage contextual cues and user behavior to adapt content delivery in real time.

One established approach to enhance visitor interaction is the deployment of humanoid robots in cultural or educational spaces. These robotic guides can attract visitors, answer questions, and provide explanations [1]. However, their high cost, maintenance requirements, and reliance on advanced onboard sensors often limit their scalability. To mitigate these constraints, AI workloads are increasingly offloaded to remote servers via fast network connections such as 5G [5]. Although this can improve responsiveness and reduce hardware demands, it still poses challenges in terms of infrastructure costs, network reliability, and latency.

The emergence of LLMs, particularly generative models such as OpenAI's GPT series, has opened new avenues for building conversational and interactive systems. LLMs offer strong capabilities for real-time question answering and natural language generation [6], making them ideal for visitor-facing applications such as chatbots. However, deploying LLMs in public settings introduces practical concerns regarding Internet dependency, response delays, and data privacy issues. A compelling solution is to integrate LLMs within an AIoT framework in which local sensors capture the user context and nearby edge servers manage language processing. Recent systems have adopted this model using vision analytics to infer demographic features and adapt content accordingly [7], [8], enabling a shift from passive displays to intelligent, context-aware digital presentations [9].

This study presents a cost-effective, AI-powered interactive presentation system that integrates AIoT sensing with RAG-augmented LLMs to deliver personalized and engaging consumer experiences. Designed as a scalable alternative to robotic guides, the Gen-Presenter uses off-the-shelf hardware and open-source models to detect presence, estimate age groups, and dynamically adjust both spoken and visual content. It supports voice-based queries, generates real-time explanations tailored to listener demographics, and delivers them through synchronized texts and audio. By combining edge computing for low-latency responsiveness with cloud-based LLMs for intelligent content generation, Gen-Presenter enables human-like interactions and adaptive storytelling, making it a compelling consumer electronics solution for museums, exhibitions, smart retail, and other public installations.

BACKGROUND AND RELATED WORKS

AI-powered multimedia presentations are becoming increasingly prevalent as public venues seek more intuitive forms of human-computer interaction. Robotic museum guides illustrate this evolution, with one system employing a 'speak-and-retreat' method to identify visitors and proactively deliver exhibit explanations [10]. A detailed review [11] acknowledges the communicative capabilities of such robots but also highlights their critical limitations, including high costs, ongoing maintenance, and safety issues in densely populated environments. To mitigate these challenges, recent strategies have shifted AI processing to remote servers via high-speed networks, such as 5G, enabling lighter and remotely managed robotic systems [5]. This distributed framework shapes the design of the Gen-Presenter, allowing for intelligent user interaction without the complexities of physical mobility requirements.

Smart digital signage and interactive kiosks are evolving beyond traditional touchscreen interfaces, increasingly incorporating gesture and voice controls to foster an intuitive user experience [5]. Modern systems can now track gaze, interpret speech, and apply computer vision techniques to infer demographic information, thereby enabling real-time content adaptation. As highlighted in [7], tailoring presentations using anonymous viewer data through tone modulation or humor integration can significantly enhance user relevance. This approach closely aligns with the Gen-Presenter's adaptive methodology.

Advancements in LLMs have introduced more nuanced conversational and generative capabilities. Although models such as GPT-3.5 and GPT-4 produce highly natural outputs, their reliance on cloud infrastructure raises concerns regarding latency, operational costs, and data privacy. In response, recent research has focused on deploying compact, edge-capable models, such as Meta's LLaMa series (7B to 13B parameters), which are suitable for execution on local servers or high-performance PCs [12]. Fine-tuned "Instruct" variants of these models enable fast, personalized interactions, making them well-suited for responsive, privacy-conscious applications such as the Gen-Presenter.

In a related initiative, [8] presented ElderEase

AR, an Augmented Reality system powered by a Multimodal LLM aimed at supporting older adults in managing daily tasks. The system enables users to capture photos and pose context-aware questions, offering tailored and real-time assistance that addresses individual needs. This project exemplifies the potential of combining AR with AI to deliver adaptive, user-specific support, particularly for seniors seeking greater autonomy. It also underscores the broader value of dynamically adjusting content based on user characteristics, which is a foundational concept within the Gen-Presenter framework.

Gen-Presenter builds on prior work by introducing audience-aware personalization using the server-hosted LLaMA 3.3B model [13], which delivers fast, cost-effective responses that are ideal for interactive settings such as museums and exhibitions. We chose LLaMA 3.3B because of its strong open-source performance, efficient multilingual inference, and suitability for edge cloud deployment. Future work will explore benchmarking Gen-Presenter against emerging LLMs, such as Mistral, MiniCPM, and Qwen2, to assess improvements in speed, model size, and language adaptability [14]. To further enhance the relevance and accuracy of its outputs, Gen-Presenter utilizes a RAG architecture that grounds its responses using semantically matched, knowledge-based entries [15]. Initially popularized by [16], RAG improves factual reliability and minimizes hallucinations in applications where precise knowledge-driven dialogues are essential.

Relying on a vector database for semantic retrieval [1], RAG has been successfully applied in domains such as TTS synthesis [17] and chatbot systems developed by companies such as NVIDIA and AWS [15]. Within the Gen-Presenter, this architecture facilitates exhibit-specific retrieval from a curated collection of slides and documents, thereby enriching both the personalization and accuracy of the content presented. In parallel, [18] demonstrated how adaptive CNNs can be used for artifact recognition and targeted content delivery, yielding substantial improvements in engagement precision. These developments reflect the expanding role of AI, particularly computer vision, in transforming static exhibitions into interactive, user-tailored experiences.

While Gen-Presenter builds upon earlier innovations aimed at enhancing visitor engagement, it uniquely addresses the challenge of adapting pre-existing content, such as presentation slides, to audience characteristics. Unlike traditional AR-based museum guides, which rely on personal devices or headsets, Gen-Presenter operates via a shared, touch-free public display that lowers participation barriers and ensures a consistent user experience. By integrating RAG, LLMs, and edge computing, this system enables real-time personalized communication. It estimates traits such as age group and monitors user attention to dynamically select relevant slides and generate context-sensitive explanations delivered simultaneously through on-screen text and synthesized speech for an immersive and adaptive experience [19].

SYSTEM ARCHITECTURE AND DESIGN

Figure. 1 illustrates the architecture of the Gen-Presenter, which comprises two main components: an edge-AI Computing Device (ECD) for on-site interaction and a back-end NLP server for computationally intensive AI tasks. This edge-cloud setup enables responsive operations while offloading resource-heavy processing, thereby optimizing both performance and cost. The ECD, a compact kiosk-style unit equipped with an Intel i7-9750H CPU, NVIDIA GTX 1660 Ti GPU, and peripherals (camera, microphone, speakers, and display), executed several AIoT modules in real time. The Gaze Detection and Age Estimation (GDAE) module uses a gaze tracker and an adapted DeepFace model [20] to assess viewer presence, engagement, and age group (child, adult, and elder), enabling age-appropriate content delivery. The Automatic Speech Recognition (ASR) module captures and transcribes speech using OpenAI's Whisper [21], offering robust multilingual transcription even in noisy environments with planned future real-world evaluations. Once the responses are generated, the TTS module uses GPT-SoVITS [22] to generate a natural voice output. Although multi-voice support is feasible for greater personalization, a generic voice was used in the prototype. The Data Processing and Control (DP) unit governs the interaction flow using state machine logic, coordinates all local modules, packages inputs (e.g., ASR transcripts, age estimates), and manages server communication and response integration.

The NLP server powers Gen-Presenter's core intelligence, hosting two LLMs (LLM1 and LLM2) and a vector database for Retrieval-Augmented Generation (RAG). Deployed on an NVIDIA DGX system with a V100 GPU [23], the server handled scalable language processing workloads. LLM1, fine-tuned for slide selection, maps user queries and demographics

This article has been accepted for publication in IEEE Consumer Electronics Magazine. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MCE.2025.3591257

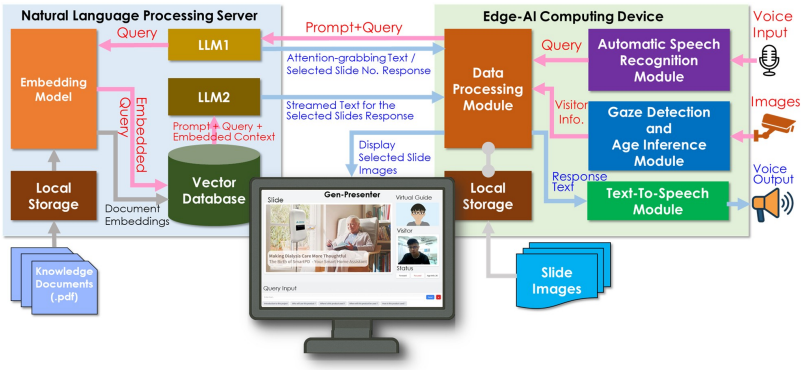

Figure 1: System Architecture of Proposed Gen-Presenter, an AIOT-based Interactive System Designed for Consumer Electronics Applications in Publication Engagement Settings.

to relevant slide identifiers and generates proactive prompts when no direct question is provided. With prompt engineering and query-type classification, the 8B model delivered near-human selection accuracy. LLM2, built on the same LLaMa 3.3 model [13], generates concise age-sensitive explanations based on the slide context, user input and real-time RAG augmentation. Its responses are streamed to the ECD as they are produced, thus minimizing latency. The vector database supports RAG by embedding exhibit materials and reference documents into a searchable format; upon query receipt, LLM2 retrieves semantically relevant segments to ground its responses, thereby reducing hallucinations and improving accuracy [15].

The ECD and NLP servers communicate through a local Ethernet or Wi-Fi connection, exchanging query data, demographic context, slide IDs, generated content and control commands. To enhance reliability, the ECD includes a fallback mode (e.g., pre-recorded presentations) if the server becomes unreachable. This hybrid architecture aligns with best practices in IoT-generative AI integration [23], [24], maintaining low-latency user interactions and advanced language processing while keeping hardware requirements and costs manageable.

Given that the Gen-Presenter system processes sensitive visual inputs through the GD-AE module, privacy compliance is essential. To comply with the GDPR and CCPA, facial data from the GD-AE module are processed entirely on the device, without storage or transmission. No personally identifiable information was retained, and visual data were discarded immediately after analysis. The system adheres to data minimization principles and performs all inferences locally to safeguard the user's privacy.

OPERATIONAL WORKFLOW OF GEN-PRESENTER

Once deployed, the Gen-Presenter follows a continuous interactive workflow, as shown in Figure 2. The system monitors the surroundings in real time and actively initiates visitor engagement when it is triggered. The typical operational cycle is summarized as follows.

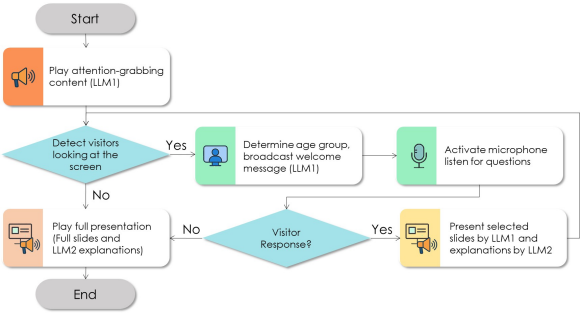

Figure 2: Operational Flow of the ECD Device in Gen-Presenter for Visitor Engagement.

- **User Detection and Input Processing:** When no active engagement is detected, ECD goes into attract mode. During this phase, LLM1 generated a short, engaging slogan or fact regarding an exhibit. This message was played via the TTS module and accompanied by a relevant visual screen to draw attention. Once the GD-AE module detects a visitor looking at the display, it estimates their age group and triggers LLM1 to generate a tailored greeting, lively and informal for children and more formal for adults. The greeting was delivered through speakers, along with on-screen prompt-encouraging questions. If the visitor responds verbally, the ASR

module captures and transcribes their query, which the DP module packages with contextual information and forwards it to the NLP server. LLM1 and LLM2 process the input to generate an appropriate, personalized response.

- **Slide Selection and Explanation Generation:** Upon receiving the visitor's query and age context, LLM1 classifies the question type (e.g., Who, What, Where) to guide slide selection. It then selects up to five relevant slides from the deck to provide a focused response to the query. Simultaneously, the query was reformulated or embedded for a semantic search against a vector database to retrieve the supporting content for grounding. LLM2 then uses the visitor's query, selected slide summaries, age context, and retrieved references to generate coherent audience-tailored explanations. Responses were structured around the selected slides and adapted in tone and complexity to match the audience, being simplified and deliberate for seniors and more casual and dynamic for teenagers. The generated text was streamed to the ECD for real-time presentation.
- **Multimodal Output Handling:** The DP module initiates multimedia presentation as soon as the LLM2 response is received. The selected slides (e.g., #2, #3, #5) are pulled from local storage and displayed sequentially, whereas the corresponding explanation segments are narrated through the TTS module. Streaming allows the narration to begin alongside slide loading, ensuring smooth audiovisual delivery without noticeable delays. After the explanation, the system briefly listens to the follow-up queries. If no further interaction occurs, it politely closes with a prompt like, "Feel free to ask another question, or I will resume the slideshow." Without renewed engagement, the system either returns to the Attract Mode, generating a fresh callout via LLM1, or plays a fallback presentation to avoid idle time.

Real-time performance is essential for maintaining smooth interactions. This setup creates a natural, dialogue-like flow: visitors are greeted, asked questions, and receive narrated visual responses quickly. In multiperson scenarios (e.g., families), responses are currently tailored to the first detected visitor, although future updates may support multi-user adaptation. The system focuses on a simple one-question, one-answer loop, avoiding complex dialogue management and fitting the typical interaction style found in exhibition environments.

PARALLEL PROCESSING PIPELINE

Figure. 3 illustrates the hybrid sequential-parallel pipeline employed by Gen-Presenter to deliver dynamic multimedia presentations. The process starts with a sequential main thread: GD-AE collects demographic data, and ASR records visitor inquiries. Once processed, LLM1 selected contextually relevant slides for the user.

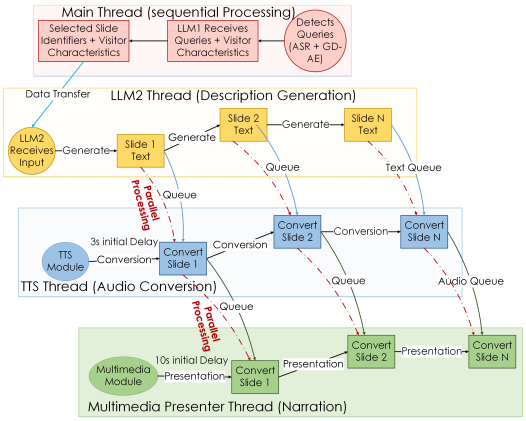

Figure 3: Parallel Processing Pipeline Showing Multi-threaded Execution with Timing Delays.

The parallel phase commences when LLM2 receives the selected slide identifiers and other metadata. To reduce latency, three threads operate concurrently: the multimedia presenter (narration), TTS, and LLM2 (description generation). The optimal throughput is achieved when these threads are synchronized using strategic delays (3s when the TTS starts and 10s when the presenter buffers). This workflow creates an assembly-line style workflow in which the presenter narrates Slide N, LLM2 generates information for Slide N+2, and TTS processes the audio for Slide N+1. Compared with a fully sequential approach, this method reduces the end-to-end latency by 63%. Empirical evaluation showed balanced CPU utilization across threads (72-78%) and a 92% user acceptance rate for seamless transitions.

The staggered delay architecture, which is carefully tuned to balance the processing overhead with natural pacing, is a significant advancement. LLM2 consistently stays ahead by one to two slides to maintain an uninterrupted flow, whereas thread-safe queues isolate processes and prevent bottlenecks. This design scales efficiently and meets the requirements of real-time AI-driven multimedia presentations. Overall,

the results highlight that well-orchestrated parallelism can deliver reliable human-like presentations without compromising the system performance.

EXPERIMENTAL RESULTS AND EVALUATION

We evaluated Gen-Presenter in three main aspects: (1) accuracy of slide selection versus human judgment, (2) suitability of the generated explanations between age groups, and (3) system responsiveness. The aim was to verify not only functional correctness but also the system's ability to mimic human-like decision-making and adjust language appropriately for different audiences. Furthermore, we compared the RAG-enhanced LLM with a baseline model to assess the improvements in output quality and latency. The system was evaluated using two versions of the Llama 3-8B model (Llama 3.2 and Llama 3.3) as LLM1 and LLM2, to assess the alignment of slide selection appropriateness with human judgment.

Test Setup

We developed a realistic demonstration scenario featuring a tech product, with a 16-slide deck and five representative visitor questions based on the 4W1H framework (”What,” ”Who,” ”Where,” ”When,” and ”How”). The two deck versions, one for children under 15 years of age and the other for adults, differed slightly in tone and focus. A technical document was embedded in the vector database to simulate background knowledge for RAG retrieval. For our evaluation, we enlisted 50 participants who were familiar with the material and served as the reviewers.

We recorded (1) LLM1-selected slides and (2) LLM2-generated RAG-tailored explanations (for both Llama 3.2-3b-instruct-fp16 and Llama 3.3:latest) for each of the five illustrative questions. To provide a ”ground truth” for comparison, ten human participants independently selected the slides that they considered suitable. In parallel, 50 evaluators assessed whether the age-specific explanations provided by the system were suitable for the target audience. We tested two system configurations: one using LLM1 alone (without RAG) and another incorporating RAG-assisted retrieval (LLM2) to support both slide selection and content generation.

Slide Selection Performance

Table 1 and Table 2 show the slide selection results for each question. For readability and layout purposes, the results are divided into two parts. Table 1 (part 1) reports the slide selections and precision, whereas Table 2 (part 2) provides the recall, accuracy, and processing times. We evaluated the performance using standard information retrieval metrics: precision (the proportion of system-selected slides matching the human choice), recall (the proportion of relevant slides correctly identified), and accuracy (overall binary correctness per slide). The processing time per query was also measured for both configurations, LLM1-only and LLM2 with RAG, in this study. The average query processing times were recorded to evaluate the efficiency of the proposed methods.

- **LLM1 (no RAG) Performance:** LLM1 (Llama 3.3) achieved strong alignment with human choices, with an average precision of 0.95, recall of 0.91, and an accuracy of 0.96 across five questions. It performed best on narrow queries (e.g., ‘ What is this product?’) often reaches 100% precision by selecting fewer highly relevant slides. Recall and Accuracy were lower for broader queries such as ”Who” and ”How” owing to their design bias towards concise, clear responses. The slide selection was fast, averaging one second per query.
- **LLM2 with RAG Performance:** When LLM2 (Llama 3.3) uses RAG, precision, recall, and accuracy dropped to 0.6, 0.35, and 0.82, respectively. Selections often include semantically related but contextually misaligned slides, reflecting a preference for textual similarity over human intent. The processing time also increased to 13.38s per query, driven by the vector search and expanded text generation. Given these trade-offs, LLM1 was preferred for slide selection, with RAG reserved to enhance LLM2's explanation quality.

For instance, when answering the question “How is this product used?” Human participants identified slides #12-16 as relevant. LLM1 (no RAG) selected # 12-15, achieving strong results (precision 0.95, recall 0.91) despite missing one slide #16). In contrast, LLM2 + RAG selected #13-15, missing critical slides (#12, #16), lowering precision to 0.6 and recall to 0.35. Consistent patterns across other queries confirmed that the two-stage approach, LLM1 for slide selection and LLM2 with RAG for explanation, outperformed the single RAG-driven model for both tasks.

Age-Appropriate Explanation Evaluation

Table 3 presents the suitability of the explanations generated by the Gen-Presenter across different age groups. Evaluators rated responses as ”suitable” or

This article has been accepted for publication in IEEE Consumer Electronics Magazine. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/MCE.2025.3591257

Table 1: Experimental Results on Slide Selection Performance (Part 1)

Queries	Slide Selection					Precision			
	HS	L3.2	LR3.2	L3.3	LR3.3	L3.2	LR3.2	L3.3	LR3.3
Q1. What is this product?	#1, #2	#1	#1	#1, #2	#2	1	1	1	1
Q2. Who will use this product?	#3, #4, #5, #6	#3, #10	#4, #8, #10	#2, #3, #5, #6	#2, #7	0.5	0.33	0.75	0
Q3. Where is this product used?	#7, #8, #9	#3, #8	#4, #5, #6	#7, #8, #9	#7, #9	0.5	0	1	1
Q4. When will this product be used?	#10, #11	#4, #10	#2, #16	#10, #11	#12, #14	0.5	0	1	0
Q5. How is this product used?	#12, #13, #14, #15, #16	#12, #15	#14, #16	#12, #13, #14, #15	#13, #14, #15	1	1	1	1
					Overall	0.7	0.46	**0.95**	**0.6**

HS: Human Selection; **L3.2**: Llama3.2; **LR3.2**: Llama 3.2 + RAG; **L3.3**: Llama 3.3; **LR3.3**: Llama 3.3 + RAG

Table 2: Experimental results on slide selection performance (Part 2 of Table 1)

Queries	Recall				Accuracy				Processing Time (s)			
	L3.2	LR3.2	L3.3	LR3.3	L3.2	LR3.2	L3.3	LR3.3	L3.2	LR3.2	L3.3	LR3.3
Q1	0.50	0.50	1.00	0.50	0.00	0.94	1.00	0.94	0.25	8.33	0.89	23.16
Q2	0.25	0.25	0.75	0.00	0.75	0.69	0.88	0.63	0.26	1.14	0.86	14.12
Q3	0.33	0.00	1.00	0.67	0.81	0.63	1.00	0.94	0.20	1.12	1.33	14.13
Q4	0.50	0.00	1.00	0.00	0.88	0.75	1.00	0.75	0.19	1.09	0.86	14.13
Q5	0.40	0.40	0.80	0.60	0.81	0.81	0.94	0.88	0.23	1.08	1.07	1.36
Overall	0.39	0.23	**0.91**	**0.35**	0.65	0.76	**0.96**	**0.82**	0.22	2.16	**1.002**	**13.38**

HS: Human Selection; **L3.2**: Llama3.2; **LR3.2**: Llama 3.2 + RAG; **L3.3**: Llama 3.3; **LR3.3**: Llama 3.3 + RAG

Table 3: Experimental Results on Slide Selection Performance by Age Group

Queries	Under 15	15-65	65+
Q1: What is this product?	0.72	0.96	0.82
Q2: Who will use this product?	0.74	0.96	0.82
Q3: Where is this product used?	0.88	1.00	0.90
Q4: When will this product be used?	0.82	0.94	0.82
Q5: How is this product used?	0.90	0.96	0.90
Overall	**0.81**	**0.96**	**0.85**

"not suitable" for children (under 15 years), adults (15-65), and seniors (65+ years), with scores reflecting the proportion of positive ratings (ideal=1.0).

Gen-Presenter showed strong adaptive capability, with all groups achieving average suitability ratings above 0.80. The highest performance was observed in the 15–65 age group, with an overall suitability of 0.96, suggesting that the system's base model aligns with adult-oriented Internet and instructional content. Across all five query types, this group's ratings remained consistent between 0.94 and 1.00.

The overall average suitability for children under 15 years of age is 0.81. Most queries scored above 0.80, with particularly strong reception for concrete, context-grounded prompts like "Where is this product used?" (0.88) and "How is this product used?" (0.90). Minor dips, such as 0.72 for "Who will use this product?" and 0.74 for "Who will use this product?" may reflect occasional vocabulary mismatches or abstraction levels that exceed younger users' comprehension level.

For seniors (65+ years), the average suitability was 0.85, indicating a robust performance. Scores remained consistent across all questions (mostly 0.82-0.90), with the highest rating for "Where is the product used?" (0.90). Feedback indicated that seniors appreciated the respectful tone and clarity, although some missed contextual anchors or modern references that could slightly affect their relatability. Overall, the results confirm that age-specific prompt tuning enables the Gen-Presenter to dynamically adjust its communicative style, emulating the natural adaptability of human guides. However, further gains, particularly for children, can be achieved using fine-tuned child-friendly educational corpora.

Single vs. Multi-thread Processing Time Performance

Figure. 4 compares the per-slide processing times of the single-threaded and multithreaded versions of the Gen-Presenter system. In the single-threaded setup, the slides are processed sequentially, including selection, description generation, and TTS conversion, resulting in cumulative latency. For 16 slides, this approach took approximately 4717s.

The multithreaded implementation parallelizes these components across the different threads. While the first slide incurs a brief delay (3s for TTS readiness and 10s before presentation), the pipeline quickly stabilizes, achieving a consistent per-slide time of roughly 18.73s from selection to output. Overall, the parallel architecture reduced the total time for 16 slides to 299.7s, a 15.7x speedup over the sequential mode, demonstrating the efficiency gains of concurrent processing.

CONCLUSION

This paper presents Gen-Presenter, a cost-effective AI-powered presentation system designed for personalized content delivery in consumer-oriented public spaces. By combining edge AI, LLMs, and a RAG framework, the system dynamically adapts slides and

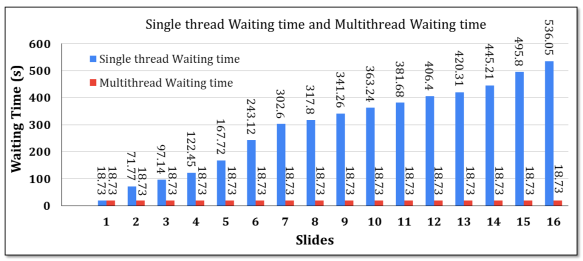

Figure 4: Slide-wise Waiting Time for Single and Multi-thread Execution.

spoken explanations in real-time based on user demographics and engagement. Experiments demonstrated that LLM1 achieved high alignment with human slide selection, whereas LLM2 with RAG effectively enriched explanations. The system demonstrated strong performance across age groups and achieved significant speed improvements through multithreaded execution. Future work will focus on enabling multi-user and multilingual capabilities, enabling Gen-Presenter to handle interactions with multiple visitors simultaneously. Additionally, with advances in model compression and edge hardware, the full inference pipeline could be deployed locally, eliminating reliance on back-end infrastructure. These enhancements position Gen-Presenter as a scalable, intelligent consumer electronics solution for public installations, with further potential in educational spaces such as classrooms and libraries, where its adaptive narration and interactive capabilities can enhance personalized learning and accessibility.

ACKNOWLEDGMENTS

”This work was supported in part by the National Science and Technology Council (NSTC), Taiwan (R.O.C), under Grant NSTC 113-2221-E-218-010-.”

REFERENCES

1. Y. Tong, H. Liu, and Z. Zhang, “Advancements in humanoid robots: A comprehensive review and future prospects,” *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 301–328, 2024.
2. J. C. de Mello, G. P. N. Secci, and P. H. Ribeiro, “Robotics and ai in museums–the future of the present,” *Robotic Systems and Applications*, vol. 4, no. 2, pp. 44–58, 2024.
3. M.-S. Wang and M.-C. Chen, “A multimedia interactive presentation system based on aiot and rag-enabled large language models,” in *2025 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2025, pp. 1–3.
4. N. Sherbina. 7 best digital signage museum examples (2025). https://www.aiscreen.io/blog/digital-signage/digital-signage-samsung/#:~:text=With%20digital%20signage%20transforming%20how,digital%20signage%20engaging%20and%20informative. Accessed: April 15, 2025.
5. S. Rosa, M. Randazzo, E. Landini, S. Bernagozzi, G. Sacco, M. Piccinino, and L. Natale, “Tour guide robot: a 5g-enabled robot museum guide,” *Frontiers in Robotics and AI*, vol. 10, p. 1323675, 2024.
6. V. Alto, *Modern Generative AI with ChatGPT and OpenAI Models: Leverage the capabilities of OpenAI’s LLM for productivity and innovation with GPT3 and GPT4*. Packt Publishing Ltd, 2023.
7. Digital signage ai – how and why chatgpt for digital signage. https://kioskindustry.org/digital-signage-ai-primer-chatgpt-how-and-why/e. Accessed: April 17, 2025.
8. T. Song, Z. Liu, R. Zhao, and J. Fu, “Elderease ar: Enhancing elderly daily living with the multimodal large language model and augmented reality,” in *Proceedings of the 2024 International Conference on Virtual Reality Technology*, 2024, pp. 60–67.
9. N. S. Amarnath and R. Nagarajan, “An intelligent retrieval augmented generation chatbot for contextually-aware conversations to guide high school students,” in *2024 4th International Conference on Sustainable Expert Systems (ICSES)*. IEEE, 2024, pp. 1393–1398.
10. T. Iio, S. Satake, T. Kanda, K. Hayashi, F. Ferreri, and N. Hagita, “Human-like guide robot that proactively explains exhibits,” *International Journal of Social Robotics*, vol. 12, pp. 549–566, 2020.
11. Y. Tong, H. Liu, and Z. Zhang, “Advancements in humanoid robots: A comprehensive review and future prospects,” *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 301–328, 2024.
12. D. Huang, Z. Hu, and Z. Wang, “Performance analysis of llama 2 among other llms,” in *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2024, pp. 1081–1085.
13. “meta-llama-3-8b instruct,” hugging face. https://huggingface.co/meta-llama/Meta-Llama-3-8B#llama-3-instruct. Accessed: April 16, 2025.
14. A. Chavan, R. Magazine, S. Kushwaha, M. Debbah, and D. Gupta, “Faster and lighter llms: A survey on current challenges and way forward,” *arXiv preprint arXiv:2402.01799*, 2024.
15. R. Merritt. “what is retrieval-augmented generation, aka rag. https://blogs.nvidia.com/blog/

what-is-retrieval-augmented-generation/. Accessed: April 17, 2025.

16. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.
17. J. Xue, Y. Deng, Y. Gao, and Y. Li, "Retrieval augmented generation in prompt-based text-to-speech synthesis with context-aware contrastive language-audio pretraining," *arXiv preprint arXiv:2406.03714*, 2024.
18. J. Wen and B. Ma, "Enhancing museum experience through deep learning and multimedia technology," *Heliyon*, vol. 10, no. 12, 2024.
19. Y. Jin, M. Ma, and Y. Liu, "Comparative study of hmd-based virtual and augmented realities for immersive museums: User acceptance, medium, and learning," *ACM Journal on Computing and Cultural Heritage*, vol. 17, no. 1, pp. 1–17, 2024.
20. M.-C. Chen, M.-S. Wang, S.-C. Wu, M.-Y. Lin, C.-Y. Chien, R.-G. Hsu, W.-J. Chang, and L.-B. Chen, "Leveraging large language models with retrieval-augmented generation for an interactive slide presentation system," in *2024 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. IEEE, 2024, pp. 1–2.
21. C. Graham and N. Roll, "Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits," *JASA Express Letters*, vol. 4, no. 2, 2024.
22. Y. Jiang, T. Wang, H. Wang, C. Gong, Q. Liu, Z. Huang, L. Wang, and J. Dang, "Expressive text-to-speech with contextual background for icagc 2024," in *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*. IEEE, 2024, pp. 611–615.
23. Volta v100 datasheet, nvidia. https://images.nvidia.com/content/technologies/volta/pdf/volta-v100-datasheet-update-us-1165301-r5.pdf. Accessed: April 21, 2025.
24. A. Wittig and M. Wittig, *Amazon Web Services in Action: An in-depth guide to AWS*. Simon and Schuster, 2023.

Ming-Shun Wang earned the B.S. (2021) and M.S. (2023) degrees in Electronic Engineering from Southern Taiwan University of Science and Technology (STUST), Tainan, Taiwan. He began his Ph.D. studies at STUST in September 2023. His current interests include large language models with retrieval-augmented generation (LLM-RAG) and AI-driven Internet of Things (IoT) systems and applications.

Showkat Ahmad Bhat received a Ph.D. degree in Communication Engineering from National Tsing Hua University (NTHU), Taiwan, in 2023. He is currently a Postdoctoral Fellow at the Center of Intelligent Healthcare, Southern Taiwan University of Science and Technology, Tainan, Taiwan. His primary research interests include Edge Computing, industrial IoT, AIoT, LPWAN (LoRaWAN/NB-IoT), precision agriculture, intelligent healthcare, Artificial Intelligence, big data, and sensor networks. He is also a member of the IEEE. Contact him at showkatbhat1994@gmail.com/showkatbhat94@stust.edu.tw.

Jian-Feng Li, received the B.S. degree in Electronic Engineering from Southern Taiwan University of Science and Technology (STUST), Tainan, Taiwan, in 2024. He began his M.S. studies at STUST in September 2024. His current interests include large language models with retrieval-augmented generation (LLM-RAG) and AI-driven Internet of Things (IoT) systems and applications.

Ming-Che Chen*, (M'19) received the B.S. and M.S. degrees in computer science from Tunghai University, Taichung, Taiwan, in 2003 and 2006, respectively, and the Ph.D. degree in computer and communication engineering from National Cheng Kung University, Tainan, Taiwan, in 2014. He has been an Associate Professor in the Department of Electronic Engineering at Southern Taiwan University of Science and Technology (EE-STUST), Tainan, Taiwan, since August 2024. From September 2020 to July 2024, he served as an Assistant Professor in the same department. Since January 2022, he has served as the Director of Artificial Intelligence over the Internet of Things (AIoT) Applied Research Center at EE-STUST. His current research interests include AIoT, industrial IoT, cloud-based application system design, wireless sensor networks and wireless communication. He is a member of IEEE. He received the Best Regional Paper Award at the IEEE ICCE 2025, the Excellent Paper Award at the IEEE GCCE 2024, and the Outstanding Paper Award from the IEEE LifeTech 2020. Please contact Jerryhata@stust. edu. tw.