Personalized Vocabulary Learning through Images: Harnessing Multimodal Large Language Models for Early Childhood Education

Ching Nam Hang+ and Sin Man Ho*
+Saint Francis University, cnhang@sfu.edu.hk; *City University of Hong Kong, sinmanho6-c@my.cityu.edu.hk

Abstract - **Early childhood is a foundational period for lifelong learning, during which vocabulary acquisition plays a crucial role in language development, cognitive growth, and social interaction. Drawing on dual coding theory, which posits that pairing words with images creates parallel mental representations that enhance memory retention, it is evident that integrating text and visuals in vocabulary learning is highly effective. However, traditional methods that rely on pre-designed flashcards often fail to accommodate the diverse learning styles and individual paces of young children, emphasizing the need for more personalized and adaptive educational materials. In this paper, we explore how multimodal large language models, guided by prompt engineering, can be harnessed to generate flashcards for vocabulary learning in early childhood settings. We illustrate how customizing different image styles to suit various developmental stages and individual preferences can enrich the personalized learning experience. This strategy enables educators and parents to produce adaptive, high-quality visual content. Our study addresses key challenges such as ensuring image appropriateness, managing cognitive load, and maintaining cultural sensitivity. Ultimately, our findings underscore the urgency of integrating generative artificial intelligence into the creation of educational content. By modernizing the way vocabulary instruction is delivered, we aim to better prepare future generations with adaptive learning resources that truly reflect the diverse needs of young learners.**

Index Terms - Early childhood education, Multimodal large language models, Prompt engineering, Vocabulary learning

INTRODUCTION

Early childhood education represents a critical period in human development, during which children acquire foundational skills that shape their lifelong learning trajectory. One of the cornerstones of early education is vocabulary learning, a process that not only builds language proficiency but also enhances cognitive development and social interaction [1]. In these formative years, children are exceptionally receptive to new words and concepts, and the integration of visual stimuli, such as flashcards that combine both images and text, can significantly amplify their learning experiences [2]. According to the dual-coding theory [3], cognition is strengthened when verbal and visual information are processed simultaneously. It suggests that when learners encounter a word alongside a corresponding image, they form parallel mental representations that enhance memory retention and recall. This cognitive synergy not only aids in understanding and internalizing new vocabulary but also contributes to the overall enrichment of early learning experiences. Traditional approaches often involve pre-designed sets of vocabulary flashcards that may not fully address the unique learning styles and pace of every child. In contrast, personalized learning strategies can tailor the educational content to match individual needs, thereby fostering a more engaging and effective learning environment. As educational paradigms evolve to prioritize personalized and adaptive learning, the necessity of integrating effective visual aids becomes ever more urgent. This evolving landscape calls for innovative solutions that empower educators to deliver customized, engaging, and cognitively robust vocabulary instruction, laying a solid foundation for lifelong learning.

Recent advancements in multimodal large language models (MLLMs) have revolutionized the landscape of artificial intelligence (AI), offering unprecedented capabilities to integrate textual and visual data [4]-[8]. Traditional neural network-based models have long been criticized for their "black-box" nature, where the internal mechanisms, how inputs are processed and outputs are generated, remain largely opaque. This opacity makes it challenging for educators and other practitioners without deep technical expertise to leverage these tools effectively, as understanding the nuances of input parameters and model behavior is not straightforward. In contrast, MLLMs, such as GPT-4 integrated with image-generation systems like DALL-E and Stable Diffusion, transform the way we think about content creation. They can seamlessly translate simple textual descriptions into high-quality visual representations, thus opening up new possibilities in various domains, including education.

Recent research on large language models (LLMs) has gained significant attention, with many findings demonstrating promising results for a wide range of applications [9]-[13]. The inherent flexibility of MLLMs allows us to generate images that not only align with specific vocabulary items but also adapt to different cultural contexts and learning environments. This convergence of natural language processing (NLP) and computer vision is particularly promising for early childhood education, where the need for engaging and easily comprehensible materials is paramount. As traditional educational resources struggle to keep pace with rapidly evolving digital ecosystems, the potential for MLLMs to bridge this gap becomes increasingly apparent. MLLMs can empower educators to design personalized learning aids that cater to the diverse needs of young learners, making the learning process more interactive and accessible. Despite these exciting advancements, the application of MLLMs in early childhood education remains underexplored, warranting a focused investigation into their efficacy and practical deployment in creating adaptive learning materials.

In this paper, we propose using MLLMs for generating images, specifically designed as flashcards, to support vocabulary learning in early childhood education through prompt engineering. Our approach not only demonstrates how to generate different types of images but also explains how these varied styles can uniquely enhance learning outcomes for young children. By offering a spectrum of image options, our method caters to diverse developmental stages and individual learning preferences, ensuring that each flashcard is tailored to the cognitive needs of the learner. Additionally, the personalized nature of our approach empowers educators and even parents, who may lack technical expertise, to generate high-quality, contextually appropriate visual aids based on the unique learning pace of the learner. This democratization of content creation bridges the gap between advanced AI capabilities and the practical requirements of early childhood pedagogy.

This paper serves as a practical guide for educators and parents alike, demonstrating how to harness MLLMs through prompt engineering to generate engaging vocabulary flashcards. Our approach leverages AI-generated visuals to complement text-based learning, making vocabulary acquisition both engaging and effective. Through our study, we demonstrate the feasibility of employing AI-driven multimodal content generation as a valuable supplement to conventional methods, offering clear insights for educators, parents, and researchers on enhancing early childhood education. Ultimately, we emphasize the urgency of adopting generative AI-driven methods to revitalize educational content creation, ensuring that the next generation of learners benefits from the most advanced and adaptive learning tools available.

RELATED WORK

Traditional machine learning and deep learning techniques have long been utilized to improve educational outcomes [14]-[16]. Recently, the advent of generative AI has opened new avenues for image generation to enhance education quality. For instance, the work in [17] proposes a hands-on workshop employing text-to-image generative AI to facilitate creative making and stimulate discourse on its integration in craft education. In [18], the authors introduce a learning workshop, where high school students use text-to-image generative AI to explore future identities while learning about its technical workings, benefits, and ethical implications. The work in [19] proposes a 6-hour learning program to teach K-12 teachers about the technical implementation, classroom applications, and ethical implications of text-to-image generation algorithms. The authors in [20] explores workshop-based graphic design education using generative AI tools like Midjourney and DALL-E, guided by the Technology Pedagogical Content Knowledge model, to teach students AI visual literacy and creative keyword generation.

Prompt engineering has emerged as a vital technique to harness the full potential of generative models for educational purposes. For example, the work in [21] formalizes text-to-image AI as a new medium for art creation and explores its potential for teaching art history, aesthetics, and technique using a dataset of 72,980 Stable Diffusion prompts. In [22], the authors explore the use of prompt engineering and text-to-image generative models in structural engineering education to visually elicit student understanding of course concepts and assess their comprehension. The work in [23] integrates generative AI, specifically Stable Diffusion, into art-focused STEAM education, where students use prompt engineering to generate images and write imaginative diaries.

METHODS

Unlike LLMs that focus solely on NLP with text inputs and outputs, MLLMs integrate advanced language processing with image generation. These models are typically built on transformer architectures and employ an encoder-decoder framework. The encoder converts a text prompt into a latent representation, while the decoder, often based on generative techniques such as diffusion models, transforms this representation into a coherent image. Trained on vast datasets of paired text and images, MLLMs learn to align specific words with corresponding visual elements, allowing them to produce detailed and contextually relevant outputs. In simple terms, they can generate images based on the descriptions provided in plain text.

Prompt engineering is the practice of crafting precise textual instructions to guide MLLMs toward producing the desired results. Although there are many techniques available in prompt engineering, our focus is on using clear and specific prompts to direct these models in generating accurate flashcards for vocabulary learning in early childhood. For instance, when creating a flashcard, a prompt might specify the use of a soft, cool background color to create a visually comfortable learning environment. Research in color psychology suggests that cool tones are generally perceived as calming and may help maintain focus, thereby enhancing the overall educational experience [24]-[25]. Therefore, careful prompt design can integrate visual elements that support learning objectives and provide a positive learning experience for young children.

In this paper, we provide a structured set of instructions to guide the MLLMs in generating flashcards for vocabulary learning. This guidance is composed of several key components:

- *Objective*: This defines the primary purpose, which is to generate flashcards that effectively support vocabulary acquisition in early childhood. It ensures that the model understands the educational intent behind the generated content.
- *Constraints*: These are the specific conditions that each flashcard must satisfy to be suitable for young learners.
- *Variables*: These include adjustable parameters that tailor the flashcards to specific educational needs. In this study, we focus on four primary variables. The first, *Target*, specifies the educational level or age group for which the flashcards are designed, ensuring that the content is age-appropriate. The second, *Theme*, outlines the overall subject matter or context of the vocabulary, guiding the conceptual framework of the flashcards. The third, *Graphic Design Style*, defines the visual aesthetic and tone, ensuring that the images are appealing and suitable for young learners. Finally, the *Number of Flashcards* variable indicates how many cards should be generated, allowing for scalability in different educational settings.

An example prompt incorporating the above-mentioned key components is provided in Table I.

RESULTS

In this section, we present a series of case studies that demonstrate the visual quality of the generated flashcards produced through our prompt engineering approach, powered by GPT-4o [8] as the underlying MLLM. By applying different themes and graphic design styles, we showcase how our methodology adapts to varied educational contexts and preferences.

I. Cartoon

Cartoon graphic design styles facilitate learning among young children by presenting educational content in an engaging and cognitively accessible manner. According to the Cognitive Theory of Multimedia Learning [26], simplified and clear visuals help reduce extraneous cognitive load, allowing young learners to focus on essential information without distraction. The inherent simplicity, vibrant colors, and exaggerated features of cartoons capture children's attention and promote sustained engagement. It is also found that when visual elements are rendered in a simplified, relatable format, such as cartoons, children are better able to form mental representations and integrate new information with existing knowledge [27]. This dual coding of verbal and visual information not only aids memory retention but also enhances comprehension, making vocabulary learning more effective. By employing a cartoon style in flashcards, educational materials can foster a positive emotional connection with learners, thereby boosting both motivation and recall.

TABLE I

PROMPT ENGINEERING FOR FLASHCARD GENERATION IN VOCABULARY LEARNING

Example Prompt
Please generate flashcards that achieve the following objective while strictly adhering to the constraints and variables listed below. Objective: Generate flashcards for vocabulary learning. Constraints: 1. Maintain a consistent style across multiple flashcards. 2. Use simple images paired with clear vocabulary display. 3. Employ a simple background that emphasizes both the image and the vocabulary. 4. Ensure color matching, large fonts, and easy readability suitable for young learners. 5. Exclude any text unrelated to the image and any imagery that does not correspond to the vocabulary term. Variables: Target: Age 4 (Early Childhood Education). Theme: Sport. Graphic Design Style: Cartoon. Number of Flashcards: 5.

In Figure I, we present generated flashcards in a cartoon style across three themes: fruit, sport, and animal, with two examples for each. The flashcards feature a cohesive design, using bright colors and clear illustrations to help children quickly associate words with images. Each card prominently features a single illustration paired with a matching vocabulary term, reinforcing learning through both visual and textual cues. The playful aesthetic, particularly evident in the animal-themed cards with expressive facial features, enhances the approachability of the material and encourages active engagement. However, some inconsistencies were observed, such as variations in text size and uneven spacing, which could affect readability and the overall uniformity of the learning experience. Despite these minor issues, the flashcards are highly effective in capturing attention and supporting vocabulary learning in early childhood.

II. Photograph

Photographic images, by depicting real objects with a high degree of detail, provide young learners with concrete visual cues that enhance vocabulary acquisition. The realism offered by photographs bridges the gap between abstract words and tangible experiences, thereby facilitating the formation of robust mental representations. The rich perceptual detail in these images enables children to make precise associations between vocabulary terms and their corresponding objects, improving both recognition and recall. By grounding vocabulary instruction in real-world visuals, photographic flashcards help children connect new words with everyday experiences, reinforcing learning in a meaningful and accessible way. It is found that realistic visual representations contribute to improved comprehension and memory retention in early childhood education [28].

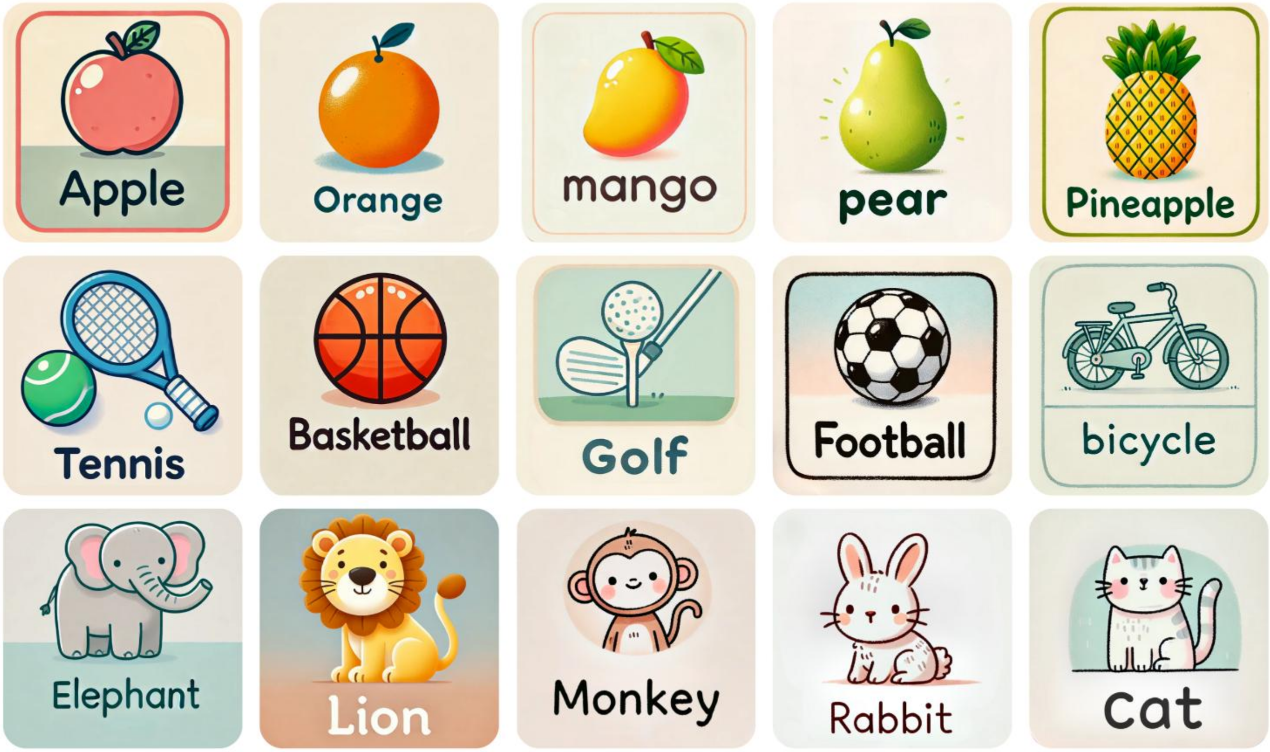

FIGURE I
GENERATED CARTOON-STYLE FLASHCARDS FOR VOCABULARY LEARNING

In Figure II, the generated flashcards in photographic style are displayed with the three themes similar to the cartoon-style flashcards in Figure I. The realistic representation of objects and animals in these flashcards provides concrete visual references that enhance vocabulary learning. The high level of detail in the photographs enables children to recognize textures, shapes, and colors as they occur in real life, making word associations more accurate and meaningful. This authenticity bridges the gap between abstract vocabulary and tangible experiences, reinforcing memory retention and comprehension. However, similar to the cartoon-style flashcards, some minor inconsistencies were observed in text size and placement, and certain images exhibit variations in lighting and background styles that could slightly detract from the overall uniformity. Despite these minor issues, the photographic flashcards remain highly effective for early learners.

Compared to cartoon-style illustrations, which use exaggerated features and bright colors to capture attention, photographic images provide a more precise and authentic representation of real-world objects. While cartoons can enhance engagement through playful aesthetics and emotional appeal, realistic images offer a direct, unfiltered connection to the real world, making them particularly useful for reinforcing practical knowledge and real-life recognition. Both styles serve valuable but distinct educational purposes, with cartoons fostering engagement and imagination, while photographic images enhance accuracy and real-world understanding.

III. Line Art

Line art is characterized by its simplicity and clarity, using clean lines and minimal detail to represent objects and concepts. This streamlined visual approach helps young learners by reducing extraneous information and focusing attention on the core features of vocabulary items. The reduced visual complexity facilitates cognitive processing, allowing children to form strong associations between words and their corresponding images, thereby enhancing memory retention [29]. Additionally, the unambiguous contours and structure inherent in line art support the integration of verbal and visual information into cohesive mental representations [30]. As a result, educational materials that employ a line art style can significantly improve vocabulary acquisition by making learning both accessible and engaging for early childhood students [28].

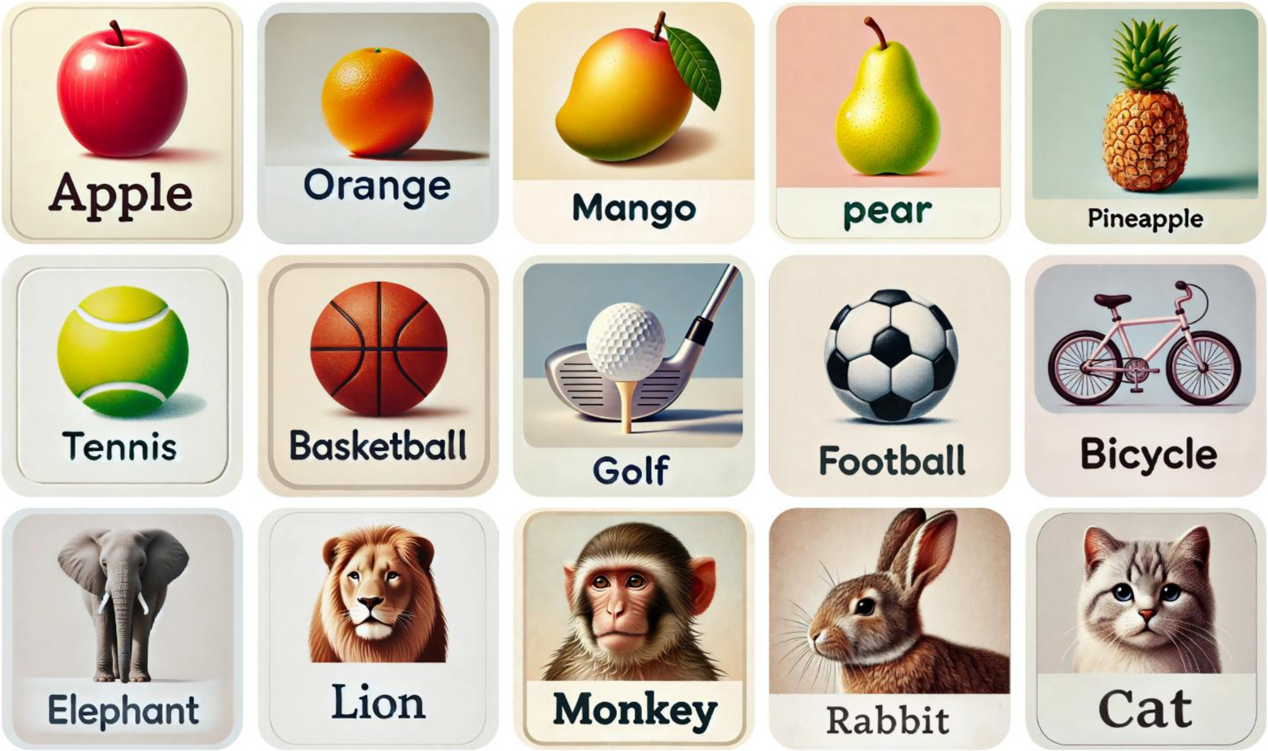

FIGURE II
GENERATED PHOTOGRAPHIC FLASHCARDS FOR VOCABULARY LEARNING

In Figure III, we observe flashcards generated using a line art style with the same three themes as seen in Figures I and II. The flashcards demonstrate that line art effectively supports vocabulary learning of children by offering clear and simplified representations of objects and animals. The minimalist design eliminates extraneous details, allowing young learners to focus on essential shapes and structures, which makes word associations more direct and less overwhelming. This approach encourages the recognition of fundamental characteristics without distractions. However, minor inconsistencies in font size and alignment are evident, and some illustrations, such as those of the lion and elephant, appear slightly more detailed than others, potentially affecting uniformity. Additionally, the absence of color may make the flashcards less visually stimulating for some children. Overall, despite these small issues, the design remains highly functional and effective for learning.

Compared to the previously analyzed cartoon-style and photographic flashcards, the line art style may not be as immediately engaging due to its lack of vibrant colors and realistic textures, yet it offers a distinct clarity that benefits vocabulary acquisition.

CONCLUSION

In conclusion, this study demonstrates the potential of leveraging multimodal large language models, guided by prompt engineering, to generate effective vocabulary flashcards for early childhood education. By experimenting with various graphic design styles, such as cartoon, photographic, and line art, we have shown that each style offers unique advantages in enhancing vocabulary acquisition. The cartoon-style flashcards, with their vibrant colors and playful aesthetics, capture young learners' attention and foster engagement, while the photographic style provides detailed and realistic images that create strong associations between words and real-world objects. Meanwhile, the minimalist approach of line art ensures clarity and reduces extraneous visual information, allowing children to focus on the core features of each vocabulary item. Despite some minor inconsistencies in text formatting and alignment across styles, the overall quality and educational effectiveness of the generated flashcards were clearly demonstrated. These findings underscore the value

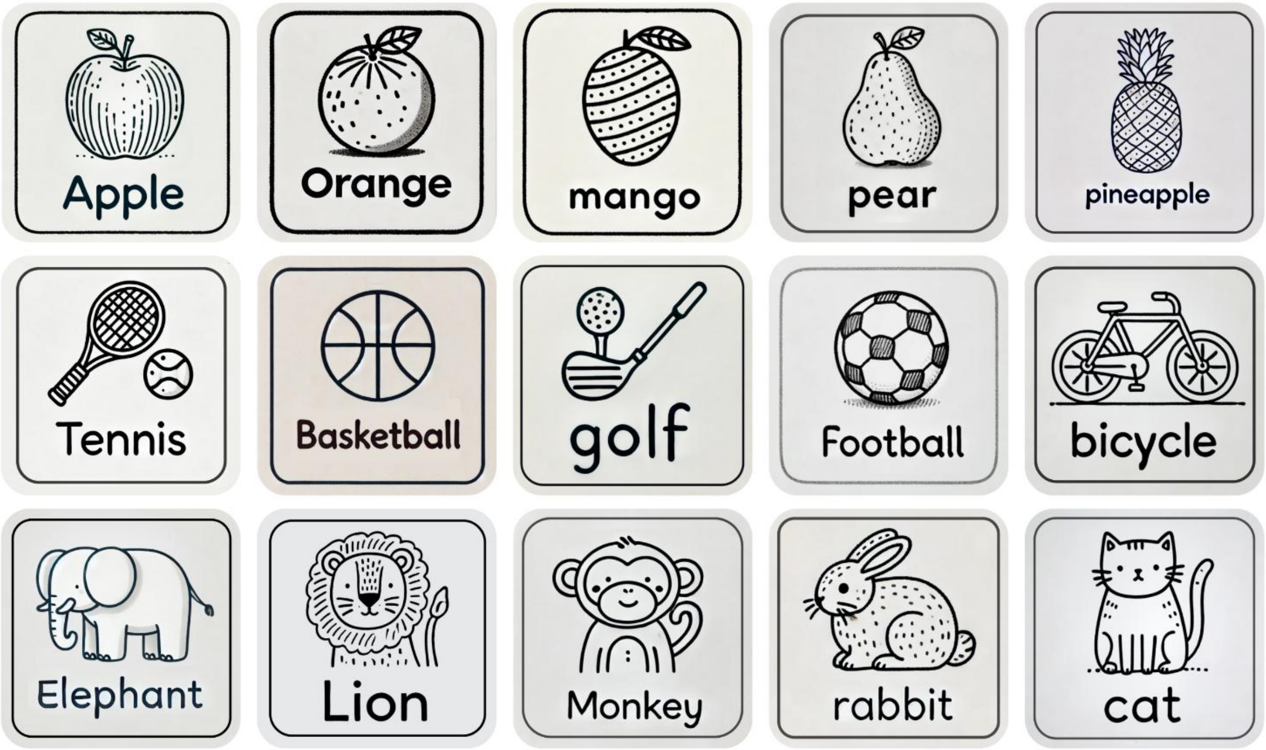

FIGURE III
GENERATED LINE ART FLASHCARDS FOR VOCABULARY LEARNING

of integrating innovative, generative AI techniques into educational content creation, ultimately empowering educators and parents to produce personalized, adaptive learning materials that meet the diverse needs of young learners. Future work will further refine prompt engineering strategies and explore additional graphic design styles to enhance both the consistency and impact of AI-generated educational resources.

ACKNOWLEDGMENT

This work was supported in part by the Saint Francis University Research Enhancement Grant (No. ISG230101-R4) and in part by the Research Grants Council of Hong Kong, under its Institutional Development Scheme Research Infrastructure Grant (No. UGC/IDS(R)11/21) and Collaborative Research Grant (No. UGC/IDS(C)11/E01/24).

REFERENCES

[1] Neuman, S.B. and Wright, T.S., 2014. “The magic of words: Teaching vocabulary in the early childhood classroom.” American Educator 38(2), pp. 4 – 13.

[2] Chen, R.W. and Chan, K.K., 2019. “Using augmented reality flashcards to learn vocabulary in early childhood education.” Journal of Educational Computing Research 57(7), pp. 1812 – 1831.

[3] Clark, J.M. and Paivio, A., 1991. “Dual coding theory and education.” Educational Psychology Review 3, pp. 149 – 210.

[4] Yin, S., Fu, C., Zhao, S. et al., 2023. “A survey on multimodal large language models.” arXiv:2306.13549.

[5] Georgiev, P., Lei, V.I., Burnell, R. et al., 2024. “Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.” arXiv:2403.05530.

[6] Wu, S., Fei, H., Qu, L. et al., 2023. “Next-GPT: Any-to-any multimodal LLM.” arXiv:2309.05519.

[7] Gong, T., Lyu, C., Zhang, S. et al., 2023. “Multimodal-GPT: A vision and language model for dialogue with humans.” arXiv:2305.04790.

[8] Achiam, J., Adler, S., Agarwal, S. et al., 2023. “GPT-4 technical report.” arXiv:2303.08774.

[9] Hang, C.N., Tan, C.W. and Yu, P.D., 2024. “MCQGen: A large language model-driven MCQ generator for personalized learning.” IEEE Access 12, pp. 102261 – 102273.

[10] Hang, C.N., Yu, P.D., Morabito, R. et al., 2024. “Large language models meet next-generation networking technologies: A review.” Future Internet 16(10), pp. 365.

[11] Hang, C.N., Yu, P.D. and Tan, C.W., 2024. “TrumorGPT: Query optimization and semantic reasoning over networks for automated fact-checking.” Proceedings of the 2024 58th Annual Conference on Information Sciences and Systems (CISS), pp. 1 – 6.

[12] Wong, M.F., Guo, S., Hang, C.N. et al., 2023. “Natural language generation and understanding of big code for AI-assisted programming: A review.” Entropy 25(6), pp. 888.

[13] Tan, C.W., Guo, S., Wong, M.F. et al., 2023. “Copilot for Xcode: Exploring AI-assisted programming by prompting cloud-based large language models.” arXiv:2307.14349.

[14] Li, J., Tan, C. W., Hang, C. N. et al., 2022. “A chatbot-server framework for scalable machine learning education through crowdsourced data.” Proceedings of the Ninth ACM Conference on Learning @ Scale, pp. 271 – 274.

[15] Tan, C. W., Ling, L., Yu, P. D. et al., 2020. “Mathematics gamification in mobile app software for personalized learning at scale.” Proceedings of the 2020 IEEE Integrated STEM Education Conference (ISEC), pp. 1 – 5.

[16] Liu, Z., Chen, J. and Luo, W., 2023. “Recent advances on deep learning based knowledge tracing.” Proceedings of the Sixteenth

ACM International Conference on Web Search and Data Mining, pp. 1295 – 1296.

[17] Vartiainen, H. and Tedre, M., 2023. “Using artificial intelligence in craft education: Crafting with text-to-image generative models.” Digital Creativity 34(1), pp. 1 – 21.

[18] Ali, S., Ravi, P., Williams, R. et al., 2024. “Constructing dreams using generative AI.” Proceedings of the AAAI Conference on Artificial Intelligence 38(21), pp. 23268 – 23275.

[19] Ali, S., Ravi, P., Moore, K. et al., 2024. “A picture is worth a thousand words: Co-designing text-to-image generation learning materials for K-12 with educators.” Proceedings of the AAAI Conference on Artificial Intelligence 38(21), pp. 23260 – 23267.

[20] Hwang, Y. and Wu, Y., 2025. “Graphic design education in the era of text-to-image generation: Transitioning to contents creator.” International Journal of Art & Design Education 44(1), pp. 239 – 253.

[21] Dehouche, N. and Dehouche, K., 2023. “What’s in a text-to-image prompt? The potential of stable diffusion in visual arts education.” Heliyon 9(6), e16757.

[22] Chacón, R., Vieira, C. and Murzi, H., 2023. “Eliciting student understanding in structural engineering classrooms using text-to-image generative models.” Proceedings of the 2023 IEEE Frontiers in Education Conference (FIE), pp. 1 – 5.

[23] Lee, U., Han, A., Lee, J. et al., 2024. “Prompt Aloud!: Incorporating image-generative AI into STEAM class with learning analytics using prompt data.” Education and Information Technologies 29(8), pp. 9575 – 9605.

[24] Elliot, A.J. and Maier, M.A., 2014. “Color psychology: Effects of perceiving color on psychological functioning in humans.” Annual Review of Psychology 65(1), pp. 95 – 120.

[25] Valdez, P. and Mehrabian, A., 1994. “Effects of color on emotions.” Journal of Experimental Psychology: General 123(4), pp. 394 – 409.

[26] Mayer, R.E. and Moreno, R., 1998. “A cognitive theory of multimedia learning: Implications for design principles.” Journal of Educational Psychology 91(2), pp. 358 – 368.

[27] Moreno, R. and Mayer, R., 2007. “Interactive multimodal learning environments: Special issue on interactive learning environments: Contemporary issues and trends.” Educational Psychology Review 19, pp. 309 – 326.

[28] Mayer, R.E., 2002. “Multimedia learning.” Psychology of Learning and Motivation 41, pp. 85 – 139.

[29] Sweller, J., 1994. “Cognitive load theory, learning difficulty, and instructional design.” Learning and Instruction 4(4), pp. 295 – 312.

[30] Paivio, A., 1990. “Mental representations: A dual coding approach.” Oxford University Press.