Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Enhancing Language Learning Through Multimodal AI-Driven Feedback on Picture Descriptions: An Eye-Tracking Study

Ruibin ZHAO[ab], Zhiwei XIE[a], Yipeng ZHUANG[a], Huixian LI[c] & Philip L. H. YU[a*]
[a]*Department of Mathematics and Information Technology,*
The Education University of Hong Kong, Hong Kong SAR, China
[b]*School of Computer Science and Information Engineering, Chuzhou University, China*
[c]*School of Smart Education, Jiangsu Normal University, China*
*plhyu@eduhk.hk

Abstract: To enhance English language learning through descriptive tasks based on everyday life scenes, we propose leveraging multimodal AI technologies. This involves utilizing multimodal large language models to automatically assess the quality of student responses and provide personalized, timely feedback in the form of AI-driven comments and suggestions for improvement. Furthermore, we will investigate students' perceptions of the feedback provided and its effectiveness in enhancing their writing skills. To achieve this, 30 participants were recruited to describe a set of daily-life pictures with the AI-driven automated assistance. During the experiment, eye-tracking technology was employed to capture students' eye movement data, enabling analysis of their visual perception and attention allocation. A survey questionnaire was administered to gather students' perceptions of this language learning approach and the effectiveness of the feedback. The experimental results indicate that students exhibit a positive attitude towards this language learning approach, as evidenced by their high levels of learning interest and motivation. Moreover, students demonstrate a willingness to actively engage with automatic feedback, with a particular inclination towards investing more attention and time in the suggestions generated by a large language model.

Keywords: Multimodal AI model, AI-driven Feedback, Eye-Tracking, Picture Description, Language Learning

1. Introduction

The rapid advancement of artificial intelligence technology has led to the development of increasingly powerful tools that facilitate second language learning. By leveraging AI, these tools provide personalized instruction, automated feedback, and immersive experiences, garnering significant attention in the field of second language (L2) learning. Researchers have developed a diverse range of tools to support various L2 learning tasks. For instance, automated writing evaluation (AWE) tools have been designed to assess the quality of student writing and offer instant diagnostic feedback (Khoii & Doroudian, 2014; Deeva et al., 2021). Moreover, AI-powered tools are available to aid pronunciation training, enabling students to practice and enhance their speaking skills through corrective feedback by analyzing the students' speech using acoustic pattern recognition and pronunciation error detection techniques (Lee, 2016). Notably, existing research has demonstrated the great potential of these AI-powered language learning tools to boost students' motivation (Wilson & Czik, 2016), attitudes toward language learning (Roscoe et al., 2018), and self-efficacy (Wilson & Roscoe, 2020). By positively impacting these important affective and cognitive factors, these tools can contribute to a more engaging and effective language learning experience for students.

Understanding language within authentic contexts is useful for effective language learning, as language is intrinsically shaped by its context and closely tied to the situations in which it is used (Lee, 2022). The opportunity to practice language in realistic and contextual

scenarios can greatly enhance learners' comprehension and practical application of the target language. In line with this, there is a growing advocacy among researchers to provide students with more opportunities to enrich their language learning experiences through engagement in authentic contexts and real-life activities (Shadiev et al., 2017; Lai, 2019; Godwin-Jones, 2022). However, research on leveraging authentic contexts to promote students' autonomous language learning remains limited. To address this gap, we propose utilizing emerging multimodal large language models to support students learning English through descriptive tasks based on everyday life scenarios. In accordance with this proposal, we have developed an online system that enables learners to observe and describe daily-life situations while receiving diverse forms of feedback from the AI-powered system. In this paper, we focus on two research questions: (1) How do students allocate their attention to the different types of feedback provided by multimodal large language models? (2) What are students' perceptions and experiences of this AI-driven language learning activity?

2. Methodology

2.1 Describing daily-life pictures with automated evaluation and feedback

Engaging children in descriptive tasks based on daily-life pictures has proven to be effective for fostering language production and developing their language skills. This approach allows children to establish meaningful connections between their language knowledge and authentic contexts, thereby promoting language learning in a more engaging and effective manner. Consequently, describing daily-life pictures has been widely adopted to support authentic language learning, as it effectively stimulates individual language production in various forms. For example, it facilitates vocabulary acquisition by requiring students to identify and name objects within a scene (Song & Ma, 2021), develops speaking and listening skills through verbal descriptions of a scene (Hwang & Chen, 2013), and enhances writing skills by encouraging students to interpret the intricate details and meaning of a whole scene (Nguyen et al., 2022). Furthermore, incorporating visual scenes and integrating them with language knowledge in these tasks aligns well with the principles of Dual Coding Theory (Clark & Paivio, 1991), which suggests that our brains process information through multiple sensory modalities. This multimodal-multisensory approach enhances knowledge understanding and facilitates memory retention, as our brains are inherently dynamic learning systems that process information in a multisensory manner.

To achieve this goal, we have developed an online system that supports students in practicing their English writing skills through descriptive tasks based on daily-life pictures. This system enables students to engage in various activities, including observing pictures, composing descriptions, receiving automated evaluation and feedback generated by AI, and continuously refining their descriptions based on the provided feedback. Our system utilizes two multimodal AI models to provide automated evaluation and feedback for students. The first model, trained explicitly for this writing task in our previous research (Zhao, et al., 2023), assesses the quality of students' descriptions by estimating the degree of alignment between the task picture and the student's description, while also analyzing the linguistic features of the description. The second model, Large Language-and-Vision Assistant (LLaVA), is a multimodal large language model capable of processing both visual and textual information and generating text content based on specified requirements (Liu, et al., 2023). In our system, LLaVA is employed to generate suggestions for improvement for students according to the content of the picture and the quality of the student's description. As a result, when a student creates a description for a given daily-life picture, the system invokes the two models to evaluate the quality of the description and generate three types of feedback information:

(1) **Scoring (Quantitative Feedback)**: Our automated scoring model evaluates the quality of the student's description by evaluating the alignment between the task picture and the description and analyzing its linguistic features. This model generates five sub-scores to evaluate the comprehensiveness, vividness, vocabulary, mechanics, and structure of the description, resulting in an overall score ranging from 0 to 10, which provides a holistic evaluation of the description’s quality.

(2) **Comments (Constructive Feedback)**. Based on the automated scoring results, our system generates a set of comments that provide detailed explanations for the scores assigned in each evaluation dimension. These comments not only highlight the strengths of the students' descriptions but also identify any weaknesses and areas that require improvement.
(3) **Suggestions (Actionable Feedback)**. In addition to the scores and comments, our system employs LLAVA to generate a set of targeted and detailed improvement suggestions for students. These suggestions take into account the content of the task picture and the students' descriptions, providing specific guidance on how to enhance their descriptions. For instance, they may remind students to include overlooked objects in the picture or provide instructions on addressing areas requiring more attention.

These feedback mechanisms empower students to progressively refine the quality of their descriptions through an iterative revision process. With each successive round of revision and evaluation, students receive updated scores and feedback, enabling them to track their progress and guiding them toward crafting more vivid and expressive language to produce improved descriptions.

2.2 Participants

Our study involved a cohort of 21 university students in China, comprising 11 females and 10 males. All participants were native Chinese speakers with over 10 years of English language learning experience. As this was a quasi-experimental study without specific controlled variables, participation was open to anyone who could coherently describe a picture without major difficulties. In accordance with ethical guidelines, all participants provided informed consent before participating in the experiment.

2.3 Experiment procedure

The data was collected from participants' engagement with feedback on their picture description tasks. Each participant was required to describe at least one picture depicting everyday objects and scenes, such as a basketball game on campus or animals in a park. For each picture, participants were instructed to provide a detailed description in English, including the characters, their actions, and the location. Additionally, when creating a description, students were encouraged to review the AI-generated feedback carefully and revise their descriptions accordingly.

The participants completed the description experiment using a laptop equipped with Tobii Studio and a portable Tobii-X3-120 eye tracker attached to the bottom frame of the laptop. Each participant completed the tasks on a scheduled day of data collection. Before the description process, the researcher conducted a 9-point calibration to calibrate the eye movements of each participant. If the calibration failed, the participant was excluded from the study. Subsequently, each participant used the online system to describe the given picture within a 20-minute time limit. They were allowed to submit multiple drafts for each picture but were prohibited from accessing external resources such as dictionaries, translation software, or their phones.

As a result, we collected 44 sets of eye gazes of viewers of video sequences, which detailed the students' gaze path and fixation duration during the experiment process. Furthermore, after the experiment, each participant completed an interview to provide insights into their learning experience, including their attitudes towards this language learning approach and their perceptions of the feedback provided by the AI system.

2.4 Areas of Interest (AOIs)

The Area of Interest (AOI) is a fundamental concept in eye-tracking data analysis, enabling researchers to isolate specific components of an interface for further analysis. By predefining AOIs, researchers can focus on particular areas or objects of interest within the corresponding stimuli, gaining a deeper understanding of visual perception, attention, and user behavior. In

alignment with our research goals, we predefined six AOIs using the supporting software for the Tobii eye-tracker to code and analyze participants' eye movement data.

As illustrated in Figure 1, our system displays task instructions in AOI-1, the picture to be described in AOI-2, and provides an input field in AOI-3 where students can enter their text descriptions. After submission, our system employs the AI models to generate scoring results, comments, and suggestions, and are then displayed in AOI-4, AOI-5, and AOI-6 respectively. This feedback supports students in revising their descriptions. It is important to note that while the categories of stimuli information displayed within these AOIs remain consistent for all participants throughout the experiment, the content of the stimulus varies, as the picture and student descriptions differ, and the AI-generated feedback content also varies depending on the picture and student descriptions.

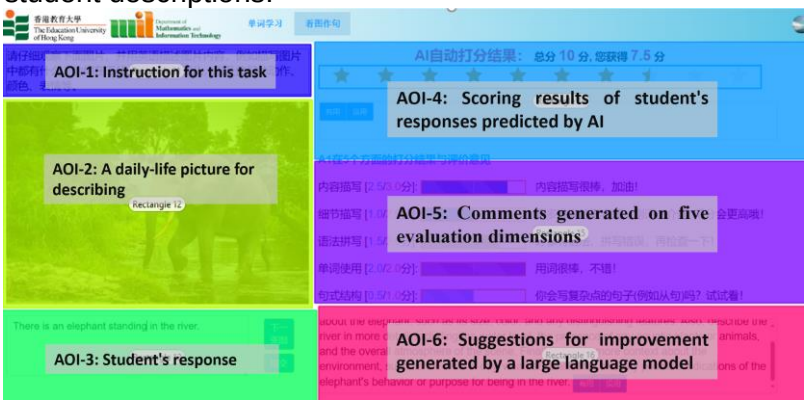


Figure 1. Six Main AOIs in This Study

3. Findings

3.1 How Students Perceive the Feedback Provided by the Multimodal AI Models

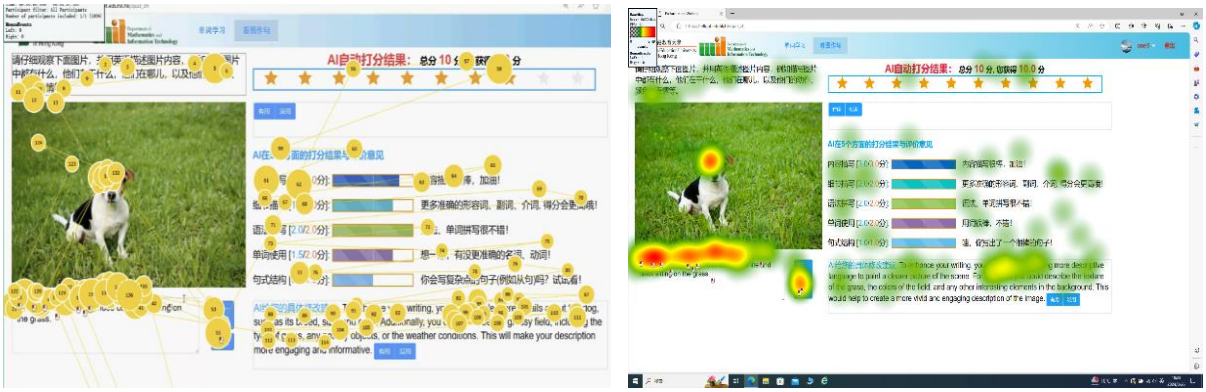
As mentioned earlier, our study utilized two multimodal AI models -- an automated scoring model and a generative large language model -- to provide students with three types of feedback: scoring results, comments on five evaluation dimensions, and suggestions for improving their descriptions. To investigate how students responded to these feedback types, particularly in terms of how they utilized the feedback to complete the picture description task, we employed eye-tracking technology to capture students' eye movement data. This technology recorded each fixation and tracked students' gaze paths, enabling us to analyze their visual attention and engagement during the task. By combining the predefined AOIs, we analyzed the number of fixations and their duration for the participants in each AOI.

Table 1. Number of fixations and their durations on each AOI

	AOI-1 Instruction	AOI-2 Picture	AOI-3 Describing	AOI-4 Scoring	AOI-5 Comment	AOI-6 Suggestion
Fixation Duration _N	8.04	100.12	368.35	23.64	56.02	152.47
Fixation Duration _Mean (Seconds)	0.13	0.16	0.18	0.14	0.15	0.16
Fixation Duration _Sum (Seconds)	1.19	16.57	66.24	3.51	8.73	24.66

The results are presented in Table 1, which shows the average number of fixations and their average duration for each AOI across the participants. Participants spent an average of 16.57 seconds observing the picture and 66.24 seconds describing it. Regarding the perception of AI-provided feedback, they allocated an average of 3.51, 8.73, and 24.66 seconds reviewing the scoring results, comments, and suggestions for improvement, respectively. This indicates that students devoted more attention to the suggestions for improvement generated by the large language model. Figure 2 provides an example

illustrating a participant's gaze path and attention hotspot, showcasing the sequence of fixations and the attention allocation when the student completed the description task.



(a) Gaze path (b) Attention hotspot

Figure 2. An example of the participant's gaze path and attention hotspot.

3.2 Student perception towards Language learning through Describing Daily Life Pictures with Automated Feedback

Following the experiment, we conducted individual interviews to gather insights into participants' learning experiences. We focused on their attitudes towards this AI-driven language learning approach and their perceptions of the AI-generated feedback. The interview results revealed that participants held overwhelmingly positive attitudes towards this innovative approach. Many students found the method engaging and enjoyable, and acknowledged the effectiveness of the AI-generated feedback. For instance, students commented, "The AI suggestions help write better descriptions", "Today's AI is amazing, it can analyze pictures", "AI reminds me to write more comprehensive and vivid descriptions".

In addition to the interviews, we administered a questionnaire featuring a set of Likert scale questions (Table 2). Each question offered five response options, ranging from 1 (Strongly Disagree) to 5 (Strongly Agree), and participants indicated their level of agreement or strength of feeling regarding each question. As shown in the second column of Table 2, the average levels of agreement for these questions, all exceeded 4.0, further confirming the participants' positive attitudes towards this learning approach and their positive perceptions of their learning experiences.

Table 2. Students' perceptions of our AI-driven language leaning system

Questions	Mean	STD
I am very satisfied with the experience of learning English through human-computer interaction.	4.20	0.68
I think these real-life scenario pictures are very helpful for learning English.	4.00	0.74
I carefully reviewed the feedback provided by AI and revised my descriptions based on it.	4.16	0.87
I agree that the feedback provided by AI can effectively help me learn English independently.	4.15	0.68

4. Conclusion

In conclusion, our study highlights the promising potential of integrating multimodal AI technologies, such as large language models, into language learning environments. By leveraging these advanced AI tools to provide personalized feedback, targeted comments, and improvement suggestions, we can create more engaging and effective student learning experiences. The experimental results reveal that participants exhibited positive attitudes towards this AI-assisted approach to learning English through picture descriptions of daily life

scenarios. Not only did they report high levels of interest and motivation, but they also actively engaged with the automated feedback, particularly valuing the suggestions generated by the large language model. This willingness to invest time and attention in AI-provided feedback underscores the perceived usefulness and relevance of such personalized guidance in supporting language skill development. The practical implications of our findings suggest that language learning practitioners and educators should explore the integration of multimodal AI technologies into their teaching practices. By harnessing the power of these advanced AI tools, they can create more personalized, engaging, and effective language learning environments, ultimately leading to better student outcomes and a more enriching language learning experience for all.

Acknowledgements

This work was supported by the CILME project fund [grant number 04A45], the research matching sub-grant [grant number CB382], a project fund under the UDSAI Research Scheme [grant number CB383], One-off Special Fund from Central and Faculty [grant number 02136] and the Start-Up Research Grant [grant number RG41/20-21R] of the Education University of Hong Kong; and the Anhui Province's University Research Projects [grant number 2023AH040216 & SK2021A0688], the Research Projects of the Chuzhou University [grant number 2022XJZD14], and the Science and Technology Plan Project in Chuzhou [grant number 2021ZD016].

References

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, 3, 149-210. https://doi.org/10.1007/BF01320076

Deeva, G., Bogdanova, D., Serral, E., Snoeck, M., & De Weerdt, J. (2021). A review of automated feedback systems for learners: Classification framework, challenges and opportunities. *Computers & Education*, 162, 104094. https://doi.org/10.1016/j.compedu.2020.104094

Godwin-Jones, R. (2022). Expanding and contextualizing digital language learning. *Bilingualism: Language and Cognition*, 25(3), 386-387. https://doi.org/10.1017/S1366728921000547

Khoii, R., & Doroudian, A. (2014). Automated scoring of EFL learners' written performance: a torture or a blessing. In *Proceedings of Conference on ICT for Language Learning* (pp. 5146–5155)

Lai, C. (2019). Learning beliefs and autonomous language learning with technology beyond the classroom. *Language Awareness*, 28(4), 291-309.

Lee, A. (2016). *Language-Independent Methods for Computer-Assisted Pronunciation Training* (Doctoral dissertation). Massachusetts Institute of Technology, Cambridge, MA.

Lee, S. M. (2022). A systematic review of context-aware technology use in foreign language learning. *Computer Assisted Language Learning*, 35(3), 294-318.

Liu, H., Li, C., Li, Y., & Lee, Y. J. (2023). Improved baselines with visual instruction tuning. *arXiv preprint* arXiv:2310.03744. https://doi.org/10.48550/arXiv.2310.03744

Nguyen, T. H., Hwang, W. Y., Pham, X. L., & Pham, T. (2022). Self-experienced storytelling in an authentic context to facilitate EFL writing. *Computer Assisted Language Learning*, 35(4), 666-695.

Roscoe, R. D., Allen, L. K., Johnson, A. C., & McNamara, D. S. (2018). Automated writing instruction and feedback: Instructional mode, attitudes, and revising. In *Proceedings of the 62nd Annual Meeting of the Human Factors and Ergonomics Society* (pp. 2089–2093).

Shadiev, R., Hwang, W. Y., & Huang, Y. M. (2017). Review of research on mobile language learning in authentic environments. *Computer Assisted Language Learning*, 30(3–4), 284–303.

Song, Y., & Ma, Q. (2021). Affordances of a mobile learner-generated tool for pupils' English as a second language vocabulary learning: An ecological perspective. *British Journal of Educational Technology*, 52(2), 858-878. https://doi.org/10.1111/bjet.13037

Wilson, J., & Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education*, 100, 94–109. https://doi.org/10.1016/j.compedu.2016.05.004

Wilson, J., & Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research*, 58(1), 87–125.

Zhao, R., Zhuang, Y., Zou, D., Xie, Q., & Yu, P.L.H. (2023). AI-assisted automated scoring of picture-cued writing tasks for language assessment. *Education and Information Technologies*, 28(6), 7031-7063. https://doi.org/10.1007/s10639-022-11473-y