LLM Enhanced Video to Text Recognition for Authentic EFL Video Making Tasks: Evaluation across GPT Versions and Video Lengths

Yi-Fan Liu [1,2], Muhammad Irfan Luthfi[3,5*], Wu-Yuin Hwang[3,4]
[1] Digital Affairs Department, Hsinchu County Government, Taiwan
[2] Research Center for Testing and Assessment, Academy for Educational Research, Taiwan
[3] Graduate Institute of Network Learning Technology, Central University, Taiwan
[4] Department of Computer Science and Information Engineering, Dong Hwa University, Taiwan
[5] Department of Electronics and Informatics Engineering Education, Universitas Negeri Yogyakarta, Indonesia
[*] iir.irfan02@gmail.com

Abstract: *Authentic video making tasks are increasingly used in English as Foreign Language (EFL) learning due to their support for meaningful communication and rich multimodal evidence, but teachers often face heavy workloads in reviewing videos and providing timely feedback. This study proposes Large Language Model (LLM)-enhanced Video to Text Recognition (VTR), which transforms videos into textual outputs, including sentence suggestions for learner narration and completes video summaries. To examine how output quality varies by video length and model version, a focused evaluation was conducted with six English teachers using a web-based system over two weeks. Participants evaluated 100 videos from a public dataset across four duration categories, rating output quality and usefulness, judging acceptability without edits, selecting error tags, and writing reference summaries for comparison. Results showed differences across Generative Pre-trained Transformer (GPT) versions and video lengths, with stronger GPT models producing higher-quality and more acceptable outputs, while longer videos led to more missing information and vagueness. These findings suggest that VTR can support EFL learning and highlight the need to improve content coverage and specificity, particularly for longer videos.*

Keywords: Video to Text Recognition, large language models, EFL learning, video summarization, authentic video making tasks

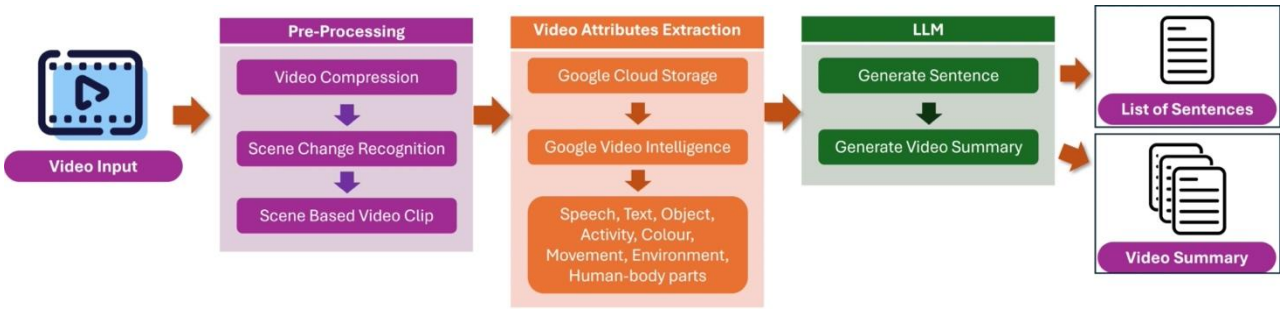
1. Introduction

Video making is increasingly used in EFL learning because it places learners in authentic situations requiring meaning planning, appropriate language use, and audience-oriented communication (Ma & Bruce, 2025), while generating rich multimodal evidence from speech, on-screen text, and visual context (Yeh, Heng, & Tseng, 2021). However, teachers often cannot review all learner videos or provide timely support (Abdelhadi & Belaid, 2025), and existing approaches relying on manual review, subtitles, or basic transcripts may fail to capture key language evidence or transform video content into reusable learning resources (Hsu, 2024), creating a gap between authentic video-making activities and scalable language support (Said Al Ghafri, 2023). To address this gap, we developed VTR, which converts video content into structured textual outputs, including sentence suggestions for learner narration and complete video summaries. Prior studies have applied VTR in authentic EFL contexts with promising results in usability, engagement, and learning support (Liu, Luthfi, & Hwang, 2024; Hwang, Luthfi, & Liu, 2024; Liu, Luthfi, & Hwang, 2025; Chang, Liu, Hwang, & Chen, 2025), but the quality of these outputs has not been systematically evaluated; therefore, this study examines the quality and perceived usefulness of VTR-generated sentence suggestions and video summaries for supporting EFL learning tasks.

2. LLM-enhanced Video to Text Recognition

Figure 1

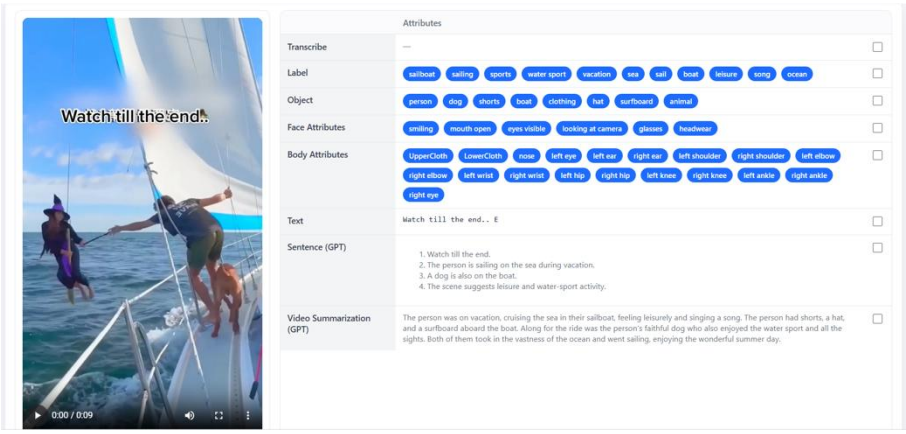
The process of LLM-enhanced Video to Text Recognition



LLM-enhanced Video to Text Recognition (VTR) transforms learner videos into structured text for language learning support. After upload (Figure 1), videos are compressed and segmented into scene-based clips using scene change recognition, then stored in Google Cloud Storage, where Google Video Intelligence extracts multimodal attributes, including speech transcription, detected text, labels, objects, activities, movement, colour, environmental information, and human body–related cues. These attributes are integrated by a large language model (GPT) to generate sentence suggestions for learner narration and a complete video summary. Figure 2 illustrates an example of a sailing video, where extracted attributes (e.g., sailboat, water sport, person, dog, and on-screen text such as "Watch till the end") are used to produce sentences describing key actions (e.g., a person sailing with a dog) and a summary of the main situation and entities, demonstrating how attribute extraction supports sentence generation and summarization for EFL learning tasks.

Figure 2

The example of VTR application to generate sentences and video summarization from a short video



3. Method

Six English teachers from a public university with EFL teaching experience participated voluntarily and provided informed consent. Data were collected over two weeks using a web-based system, in which participants evaluated system-generated sentence suggestions and video summaries, completed rating forms and a perception questionnaire, and assessed four GPT model outputs for the same 100 videos (within-rater design; 400 conditions per rater). The videos were sampled from a publicly available dataset (Shang et al., 2025) and grouped into four duration categories: very short (≤10s), short (11–30s), medium (31–60s), and long (>60s), with 25 videos per category covering diverse real-world scenes. The study evaluated two outputs: sentence suggestions (relevance, linguistic quality, level appropriateness, actionability, overall score, and error tags) and video summaries (faithfulness, coverage, coherence, conciseness, language quality, overall score, and acceptability), and participants also wrote reference summaries for ROUGE-L and BERTScore comparisons. Four model conditions were tested: GPT-3 (text-davinci-003), GPT-3.5, GPT-4, and GPT-4o. Inter-rater reliability was high (ICC = 0.87), and given the small sample size (n = 6), analysis focused on descriptive statistics and pattern comparison across video length and model conditions, with error tags analyzed using frequency counts, summary

quality examined using ROUGE-L and BERTScore against reference summaries, and associations between overall ratings and quality dimensions also explored.

4. Result and Discussion

4.1. Analysis of Quality of VTR Sentence Suggestions for Learner Narration

Sentence suggestion quality showed consistent patterns across GPT versions and video lengths, with GPT-4 and GPT-4o achieving the highest scores (approximately 4.0–4.3), followed by GPT-3.5 (approximately 3.1–3.6), and GPT-3 the lowest (approximately 2.4–3.1). Scores decreased as video length increased, with lower ratings for medium and long videos, indicating greater difficulty in generating effective narration for more complex content; across dimensions, relevance and linguistic quality were generally higher, while actionability was slightly lower, particularly for longer videos. Error tag patterns further clarified these limitations, with missing information and vagueness as the most frequent issues, especially in longer videos and weaker models, suggesting that the main challenge lies in content selection and completeness rather than grammatical correctness. When key actions were omitted or descriptions were too general, usefulness for narration decreased, while wrong content appeared more often in weaker models, indicating weaker grounding between extracted attributes and generated sentences; additionally, overly complex language and grammatical issues negatively affected level appropriateness and linguistic quality. Overall, these findings suggest that stronger GPT models provide more reliable narration support, although improvements are needed to enhance content coverage and specificity for longer and more complex videos.

4.2. Analysis of Quality of VTR Video Summary

Video summary quality and automated evaluation results showed consistent patterns across GPT versions and video lengths, with GPT-4 and GPT-4o achieving the highest performance across quality dimensions (approximately 3.8–4.3) and similarity metrics (ROUGE-L up to ~0.61; BERTScore up to ~0.91), followed by GPT-3.5, while GPT-3 showed the lowest performance. This pattern was stable across video categories, indicating that stronger models produce more accurate, coherent, and consistent summaries, whereas both quality ratings and acceptance decreased as video length increased, with the lowest performance in long videos, reflecting the difficulty of maintaining coverage and consistency in more complex inputs. Automated metrics aligned with human evaluation, showing higher similarity for stronger models and lower similarity for longer videos; among dimensions, faithfulness and coherence were generally higher, while coverage declined more noticeably with increased video length, indicating challenges in capturing key events. Overall, the results suggest that model capability primarily affects content accuracy and coherence, while video length influences completeness and information loss, highlighting the need to improve coverage and consistency for longer and more complex videos.

4.3. Analysis of Perceptions Toward VTR

Participants reported positive perceptions of VTR for supporting EFL learning tasks. Perceived usefulness was high across all items (means > 4.2), indicating that VTR can enhance task performance, productivity, and evaluation by providing effective sentence suggestions and video summaries. Perceived ease of use was also positive (means > 4.0), suggesting that the system is easy to learn, operate, and understand with minimal effort. Attitude toward use and behavioral intention were similarly strong (means > 4.2), indicating that participants viewed VTR as a valuable tool and were willing to use and recommend it in future practice. Overall, responses were consistently high, reflecting strong acceptance of VTR as a practical support tool for EFL video-making tasks.

5. Conclusion

Overall, the results suggest that VTR provides useful sentence suggestions and video summaries to support EFL video-making tasks, with positive perceptions of usefulness and ease of use. Output quality varied by model and video length, with stronger GPT models producing higher-quality and more acceptable outputs, while longer videos remained more challenging and prone to missing or vague information. This study has several limitations, including a small sample size

and the use of a fixed dataset, which may limit generalizability. Future work should involve more diverse participants and real classroom settings, examine learning outcomes, and improve the VTR pipeline to enhance content coverage, specificity, and consistency, particularly for longer videos.

References

Abdelhadi, A. & Belaid, L. (2025). Teachers' Perceptions of Implementing Student-Generated Videos as an Authentic Assessment Tool in English Language Learning. Journal of Education, Society & Multiculturalism, 6(1), 2025. 81-101. https://doi.org/10.2478/jesm-2025-0006

Al Ghafri, M. S. (2023). The effect of employing authentic video projects on EFL (English as a foreign language) students' motivation to learn English. *European Economic Letters, 13*(3), 1327–1332. https://doi.org/10.52783/eel.v13i3.435

Chang, H.-W., Liu, Y.-F., Hwang, W.-Y., & Chen, W.-C. (2025). Developing an AI driven contextualized short video learning system for EFL speaking and writing. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT 2025),* (pp. 148–152). IEEE. https://doi.org/10.1109/ICALT64023.2025.00048

Hsu, H. (2024). Evaluating the Impact of Video-Making Projects on English for Specific Purposes Learning: A Case Study. *International Journal of Language and Education Research, 6*(1), 1-23. https://doi.org/10.29329/ijler.2024.661.1

Hwang, W.-Y., Luthfi, M. I., & Liu, Y.-F. (2024). AI enhanced video drama making for improving writing and speaking skills of students learning English as a foreign language. *Innovation in Language Learning and Teaching, 18*, 1–18. https://doi.org/10.1080/17501229.2024.2437655

Liu, Y.-F., Luthfi, M. I., & Hwang, W.-Y. (2024). Developing an AI enhanced video drama making learning system to support EFL learners in authentic contexts. In *Proceedings of the Global Chinese Conference on Computers in Education (GCCCE 2024): Main Conference (English Papers)* (pp. 34–37).

Liu, Y.-F., Luthfi, M. I., & Hwang, W.-Y. (2025). Enhancing EFL writing through AI driven video to text recognition in authentic learning contexts. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies (ICALT 2025)* (pp. 166–168). IEEE. https://doi.org/10.1109/ICALT64023.2025.00053

Liu, Y.-F., Luthfi, M. I., & Hwang, W.-Y. (2025). A pilot study on bridging EFL writing and speaking skills through AI enhanced authentic short video making. In *Proceedings of the Global Chinese Conference on Computers in Education (GCCCE 2025)* (pp. 62–69).

Ma, Q. M. & Bruce, D. L. (2025). Integrating Digital Video Composing Across Three Settings in Chinese English Classrooms. In J. Tussey & L. Haas (Eds.), *Supporting Linguistic Differences Through Literacy Education* (pp. 325-360). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-3496-7.ch010

Shang, Y., Gao, C., Li, N., & Li, Y. (2025). A large-scale dataset with behavior, attributes, and content of mobile short-video platform. In *Companion Proceedings of the ACM on Web Conference 2025 (WWW '25)* (pp. 793–796). Association for Computing Machinery. https://doi.org/10.1145/3701716.3715296

Yeh, H.-C., Heng, L., & Tseng, S.-S. (2021). Exploring the impact of video making on students' writing skills. *Journal of Research on Technology in Education, 53*(4), 421–437. https://doi.org/10.1080/15391523.2020.1795955