Speech-Content-Driven Highlighting of Translated Lecture Slides for Foreign Language Lecture Understanding

Naoki Muto*, Chee Siang Leow*, Junichi Hoshino†, Takehito Utsuro†, and Hiromitsu Nishizaki*
* University of Yamanashi, Japan
E-mail: naoki_m@alps-lab.org, {leow,hnishi}@yamanashi.ac.jp Tel/Fax: +81-552208361
† University of Tsukuba , Japan
E-mail: jhoshino@esys.tsukuba.ac.jp, utsuro@iit.tsukuba.ac.jp

***Abstract*—This paper presents a lecture support system to assist international students in understanding lectures delivered in a non-native language. The system integrates speech recognition, character recognition (OCR), machine translation, and content-based highlighting. It transcribes speech using "Whisper," translates it with "DeepL," and extracts slide text and OCR using "YomiToku." "Sentence Transformers" computes similarities between spoken content and slide text, highlighting relevant translated sections. An evaluation involving 19 international students compared three system configurations: highlighting only, slide translation only, and the proposed integrated system. The evaluation results showed that the integrated approach significantly outperformed both single-feature configurations, with statistically significant differences in pairwise comparisons ($p < 0.05$) and score improvements of 20-46 points. The integrated system particularly benefited students with lower Japanese proficiency, demonstrating substantial improvements in their ability to follow the relationship between spoken content and slide information.**

I. Introduction

The internationalization of higher education has created a pressing need for effective multilingual lecture support systems. In Japanese universities, while over 270,000 international students are enrolled as of 2023[1], many face significant challenges in understanding Japanese lectures. Although universities increasingly offer English-taught courses, maintaining Japanese as the primary language of instruction remains crucial for educational quality and domestic student engagement. This situation creates a critical need for technology that enables students to understand lectures regardless of their language proficiency.

These linguistic challenges manifest in three critical ways during lectures. First, students must process specialized academic content while simultaneously struggling with basic language comprehension, creating a dual cognitive burden. Second, the dynamic nature of lectures, where instructors move between verbal explanations and visual materials, makes it difficult for non-native speakers to maintain coherent understanding. Third, the mental effort required for constant translation often prevents students from engaging deeply with the subject matter, potentially impacting their academic performance and participation. These challenges are particularly acute in science, technology, engineering, and mathematics (STEM) fields, where complex technical terminology compounds the language barrier.

Recent advances in AI technology[2]–[12] have led to the development of various AI applications. Existing commercial solutions partially address these challenges. Google Translate [13] offers image-based text translation for slides, while Microsoft PowerPoint Live [14] provides real-time slide translation during presentations. However, these tools operate in isolation, focusing solely on content translation without connecting it to the instructor's verbal explanations. Academic research has also explored various approaches: Sudo et al. [15] developed an automated subtitle generation system for lecture videos, while Bérard et al. [16] proposed an end-to-end speech translation system. More recently, Anderer et al. [17] released *"MaViLS"*, the first public benchmark for video-to-slide alignment together with a multimodal baseline, underscoring the growing interest in synchronising spoken lectures and slide content. Yet, these solutions typically handle either speech translation or slide content translation separately, requiring students to manually integrate information from multiple sources. Recently, Hotta et al. improved speech translation subtitle quality by incorporating unnecessary word detection in the automatic speech recognition (ASR) process [18], but their approach also focused solely on speech content without addressing the integration with visual materials.

This paper introduces a new lecture support system that fundamentally transforms how international students can engage with foreign language lectures. Our system uniquely integrates three key AI technologies: Whisper's robust multilingual ASR [19], DeepL's context-aware translation [20], and Sentence Transformers' semantic similarity analysis [21]. The distinguishing feature of this system lies in its ability to analyze the instructor's speech and automatically highlight slide text that corresponds to the spoken content. Without such synchronization, it is challenging for students to identify which portion of the slide the instructor is currently explaining when subtitles are displayed. However, by synchronizing the instructor's speech with dynamically highlighted text, international students can better understand the lecture content. This synchronization is achieved through a semantic matching

979-8-3315-7206-8/25/$31.00

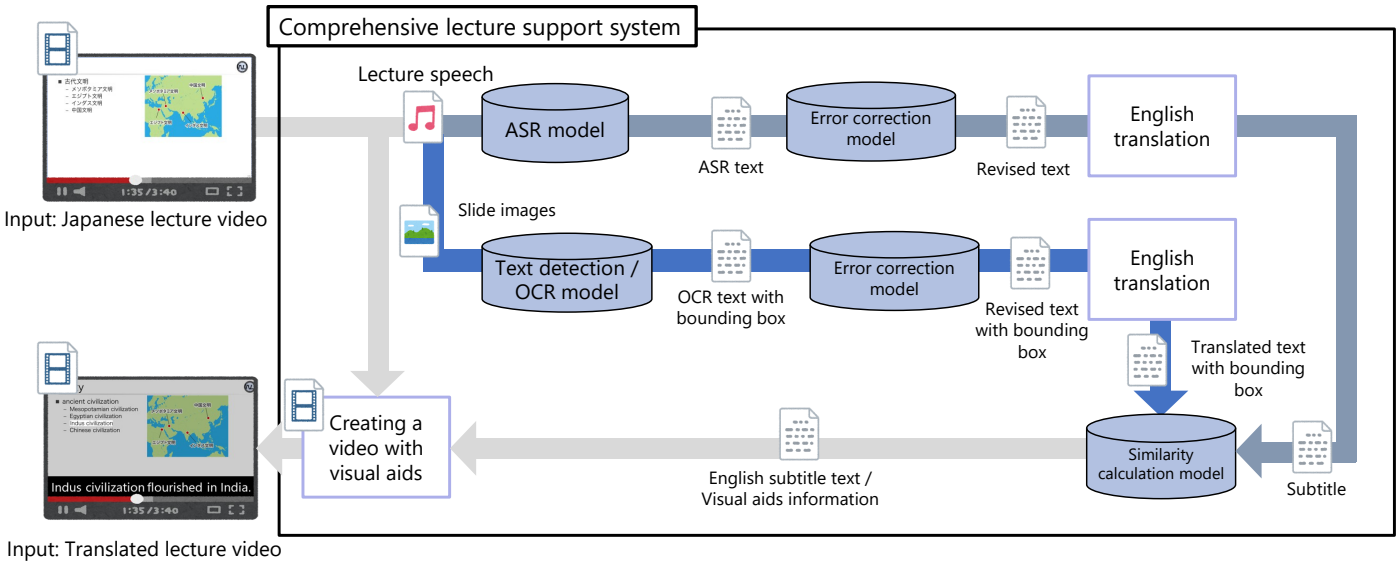

Fig. 1. Overview of the proposed multilingual lecture support system. The system integrates speech recognition, slide text extraction, and semantic matching to generate translated lecture videos with synchronized highlighting of relevant slide content.

algorithm that identifies relationships between spoken content and slide text, regardless of the source or target languages.

Our lecture support system extends beyond existing document translation systems in several key aspects. While existing solutions such as Google Translate [13] and PowerPoint Live [14] focus on translating static content, our system addresses dynamically changing content that requires continuous change detection and contextual understanding. The system employs advanced semantic similarity calculations based on deep learning embedding vectors, considering not only lexical matches but also conceptual relationships. This approach enables students to follow complex academic discussions even when specific terminology varies between languages. Furthermore, the system preserves the original slide layout by re-inserting each translated text segment into its original bounding box, keeping the visual structure of Japanese lecture slides intact and improving the reliability of downstream semantic matching.

Through a comprehensive evaluation involving 19 international students viewing translated Japanese lectures, we compared three system configurations: highlighting only, full slide translation only, and our integrated approach. The results demonstrated that our integrated system significantly outperformed both single-feature approaches and existing commercial solutions. Statistical analysis of the pairwise comparisons confirmed significant differences in preference scores between the systems. Particularly strong preferences were observed among students with Japanese proficiency below the Japanese-Language Proficiency Test [22] (JLPT)-N3 level[1], where the integrated system received notably higher selection counts. The evaluation data showed statistically significant benefits across all proficiency levels, with participants reporting improved ability to follow the relationship between spoken content and slide information in their native language. These findings demonstrate that the integration of speech analysis with synchronized visual highlighting substantially enhances lecture comprehension for international students.

The primary contributions of this research include:

- Development of a comprehensive lecture support system that integrates automatic speech recognition, machine translation, and content-based highlighting,
- Introducing a matching mechanism that continuously aligns spoken utterances with slide fragments across languages and layouts, enabling visual cues without manual intervention.
- Highlighting the corresponding slide segments while automatically translating Japanese lectures into English removes language barriers, making key points easier to understand.

II. SPEECH-CONTENT-DRIVEN LECTURE SUPPORT SYSTEM

A. System Overview

Our proposed system processes lecture videos to create a synchronized multilingual learning environment. As illustrated in Fig. 1, the processing pipeline integrates three key components: speech processing, slide content analysis, and semantic matching. The speech processing component converts lecture audio into translated subtitles through automated speech recognition and machine translation. Concurrently, the slide processing component extracts and translates text while preserving the original layout information. The semantic matching component analyzes relationships between spoken content and slide elements, enabling dynamic highlighting of relevant content. This architecture ensures synchronized

[1]The JLPT is an international standardized test for Japanese language proficiency with five levels (N1-N5, where N1 is most advanced). N3 represents intermediate-level proficiency in understanding everyday Japanese.

delivery of translated subtitles and visual guidance, creating a comprehensive learning support environment.

B. Processing Pipeline

The core processing pipeline consists of three main components:

- Speech processing: Converting lecture audio to translated subtitles
- Slide processing: Recognizing and translating slide text
- Semantic matching: Connecting speech content with relevant slide elements

a) Implementation Language: All components are implemented in Python, and the system is designed to allow smooth integration between modules such as speech, slide, and semantic processing.

1) Speech Processing: The speech processing pipeline employs Whisper (large-v3) for speech recognition, followed by ChatGPT-based error correction specifically targeting academic terminology [23]. The corrected text is then translated using DeepL API [24]. This process ensures accurate translation of technical terms while maintaining natural language flow.

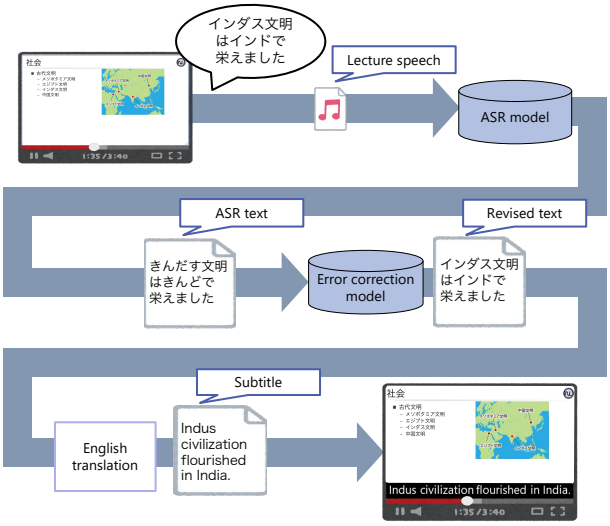

Fig. 2. Speech-to-subtitle processing pipeline. Japanese lecture speech is transcribed using "Whisper," corrected for academic terminology via "ChatGPT," and translated into English using DeepL API.

2) Slide Processing: For slide content, we first extract frames using OpenCV's frame difference analysis to identify slide transitions [25]. YomiToku [26] processes these frames to recognize text in both horizontal and vertical layouts. The recognized text undergoes ChatGPT-based error correction and DeepL translation, maintaining the original slide layout while accommodating English text length variations.

3) Semantic Matching: The system uses the BAAI/bge-m3 [27] model through Sentence Transformers to compute similarity scores between current speech and slide text elements. This matching occurs after translation to ensure consistent semantic comparison regardless of source language.

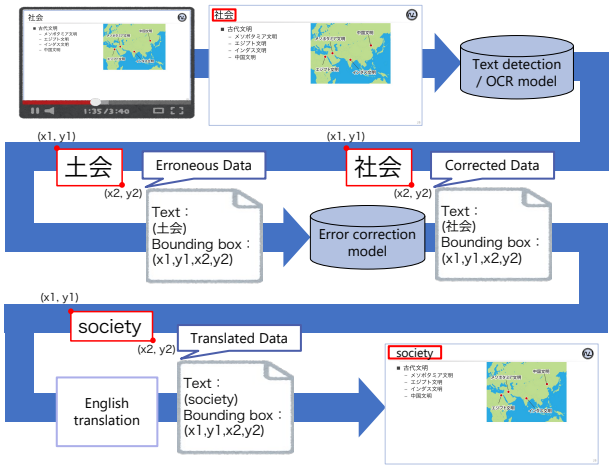

Fig. 3. Slide text extraction and translation pipeline. Japanese slide text is detected and extracted using "YomiToku OCR," corrected for recognition errors via "ChatGPT," and translated into English using "DeepL" API while preserving bounding box information.

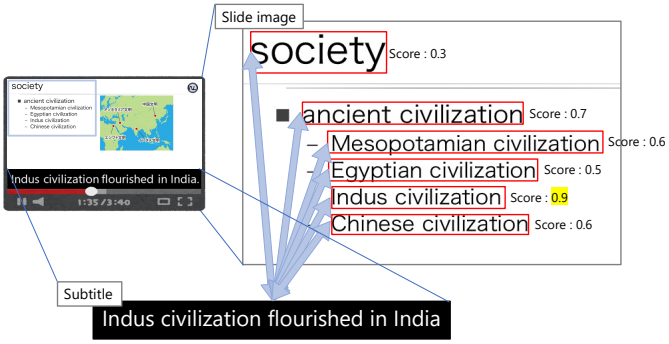

Fig. 4. Semantic similarity computation for content-based highlighting. The system calculates cosine similarity scores between the current speech segment (subtitle) and each text element on the slide using BAAI/bge-m3 embeddings to identify and highlight the most relevant content (score: 0.9 in this example).

C. System Variants

In order to compare the effectiveness of the highlights in this study, the system is divided into System 1, System 2, and System 3. Examples of output from each system are shown in Fig. 5.

1) System 1: Highlighting Only: This configuration focuses on visual guidance by highlighting slide elements that correspond to the current speech content which is shown in top-right of Fig. 5. Only the relevant portions of the slide are translated and highlighted, helping students focus on key information while maintaining the original slide layout. The highlighting mechanism includes:

- Background color enhancement for matched content
- Dynamic font size adjustment for emphasis
- Translation of only the highlighted portions

2) System 2: Full Translation: This variant provides complete translation of all slide content without highlighting which is shown in bottom-left of Fig. 5. Features include:

- Complete slide content translation

- Adaptive layout adjustment for translated text
- Preservation of original slide formatting
- Synchronized subtitle display

3) System 3: Integrated Approach: The integrated system (Proposed Method) which is shown in bottom-right of Fig. 5 combines the benefits of both approaches:

- Full slide content translation
- Dynamic highlighting of relevant content
- Synchronized subtitles with speech
- Enhanced visual cues for content relationships

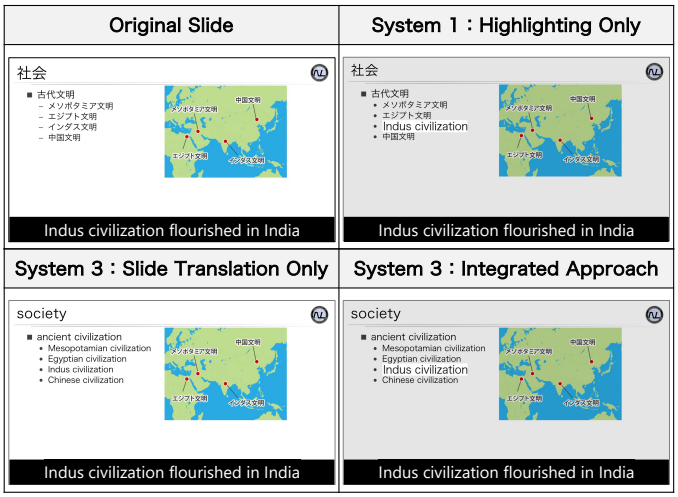

Fig. 5. Comparison of three system configurations applied to a Japanese lecture slide. From the original slide (top-left), three variants are shown: System 1 provides highlighting with partial translation (top-right), System 2 offers complete slide translation without highlighting (bottom-left), and System 3 integrates both features (bottom-right, proposed method).

III. Experimental Evaluation

A. Experiment Overview

To evaluate the effectiveness of our proposed system, we conducted a comparative study with 19 international students at Yamanashi University. The participants included students with varying levels of Japanese language proficiency, with 12 participants below JLPT N3 level.

B. Experimental Design

1) Task flow and assignment: Each participant watched 36 short lecture videos in total. Twelve different "source lectures" (topics listed in Table I) were rendered in the three system variants, giving $12 \times 3 = 36$ video files:

- System 1 (highlighting only): 12 videos
- System 2 (full translation): 12 videos
- System 3 (integrated): 12 videos

Videos were shown in pairs of semantically related topics (A/B). For each pair, the two videos were assigned to *different* systems so that participants always compared distinct systems on comparable content (e.g., "Himeji Castle" in System 1 vs. "Shirakawa-go" in System 2).

TABLE I
VIDEO PAIRS AND TOPICS

Pair	Topic A	Topic B
1	Himeji Castle	Shirakawa-go
2	Hokkaido Cuisine	Okinawa Cuisine
3	Women's Kimono	Men's Kimono
...	...	...

TABLE II
SYSTEM SCORING METHOD

Comparison result	Points awarded
Selected as more comprehensible	1 point
Not selected	0 points

2) Pairwise comparison schedule: For every participant we conducted:

- 6 comparisons of System 1 vs. System 2
- 6 comparisons of System 1 vs. System 3
- 6 comparisons of System 2 vs. System 3

Thus each participant performed $6 \times 3 = 18$ preference decisions over the 36 viewed videos. Across 19 participants this yields $18 \times 19 = 342$ total comparison trials, matching the grand totals reported in Table III.

3) Test Materials: Although public datasets such as MaViLS [17] exist for English video–slide alignment, they do not contain Japanese lectures or translated subtitles. Therefore, we produced 36 short Japanese lecture videos, each approximately one minute long, covering diverse topics. These videos are organized into pairs based on similar content and slide layouts. Table I provides an overview of the video pairs along with their respective topics.

4) Evaluation method: Each system configuration was evaluated through pairwise comparisons, with participants selecting which version provided better lecture comprehension. Table II outlines the scoring mechanism used in the pairwise comparisons of the system configurations. In each evaluation, if a participant judged one system to offer superior lecture comprehension, that system was awarded 1 point while the other received 0 points. This straightforward method enabled us to quantitatively assess participants' preferences.

C. Results

1) Overall Results: Table III presents the number of times each system was selected in pairwise comparisons, aggregated across all 19 participants. For example, "System 1 vs System 2" shows how many times participants preferred System 1 over System 2, and vice versa.

Table IV summarizes the statistical significance of the differences in system preference. The "Score Difference" column indicates the margin by which one system outperformed the other, while the p-value shows whether this difference is statistically significant at conventional thresholds (e.g., 0.05).

2) Results by Language Proficiency: Table V focuses on the 12 participants whose Japanese proficiency was below JLPT N3. It shows how many times each system was favored in

TABLE III
EVALUATION RESULTS MATRIX (ALL 19 PARTICIPANTS)

	System 1	System 2	System 3
System 1	–	47	43
System 2	67	–	34
System 3	71	80	–

TABLE IV
STATISTICAL ANALYSIS OF SYSTEM PREFERENCE

Comparison	Score difference	p-value	Significant
System 1 vs 2	20	0.047	Yes
System 1 vs 3	28	0.012	Yes
System 2 vs 3	46	0.001	Yes

pairwise comparisons within this subgroup, highlighting the pronounced preference for the integrated system (System 3).

The experimental results showed that System 3 (integrated approach) consistently outperformed both System 1 (highlighting only) and System 2 (full translation only). This preference was particularly pronounced among participants with lower Japanese proficiency levels.

D. Discussion

The experimental results provide strong evidence for the effectiveness of our integrated approach to lecture comprehension support. Table III shows clear patterns in the pairwise comparisons across all 19 participants. The proposed integrated system outperformed both the highlighting-only and full translation approaches, receiving 71 and 80 selections respectively when compared against System 1 and System 2. Statistical analysis in Table IV confirms these differences are significant, with p-values below the conventional 0.05 threshold across all comparisons. The score differences show an increasing trend: 20 points between Systems 1 and 2, 28 points between Systems 1 and 3, and the largest gap of 46 points between Systems 2 and 3.

The system's effectiveness becomes particularly evident when examining the results for participants below JLPT N3 level. Among these 12 participants, Table V shows that System 2 was notably preferred over System 1 (49 vs 23 selections), indicating that full translation provides substantial benefits for lower proficiency learners. However, the integrated system received even stronger support, with 52 and 55 selections when compared against Systems 1 and 2 respectively. This pattern suggests that while full translation alone offers significant advantages for lower proficiency learners, the addition of content-driven highlighting further enhances lecture comprehension.

TABLE V
RESULTS FOR PARTICIPANTS BELOW JLPT N3 (12 PARTICIPANTS)

	System 1	System 2	System 3
System 1	–	23	20
System 2	49	–	17
System 3	52	55	–

IV. CONCLUSION

This paper presented a new lecture comprehension support system that integrates speech recognition, character recognition, machine translation, and content-based highlighting to assist international students in understanding foreign language lectures. Our system creates a synchronized multilingual learning environment by processing lecture videos through three key components: speech processing, slide content analysis, and semantic matching. The system's key feature lies in its dynamic highlighting of relevant translated content based on speech analysis.

Through empirical evaluation with 19 international students, we demonstrated that our integrated approach significantly outperformed single-feature approaches, with statistically significant differences in pairwise comparisons ($p < 0.05$). For students below JLPT N3 level, the system showed particularly strong benefits in comprehension and engagement.

Our next step focuses on adapting the system for real-time classroom use (e.g., [28]). This requires reducing the current processing latency and optimizing the system architecture for live processing of lecture content. Future work will investigate efficient processing methods and improved semantic matching algorithms to enable real-time support during live lectures, making foreign language education more accessible to international students.

V. ACKNOWLEDGMENT

This research was supported by JSPS Grant-in-Aid JP25H00566.

REFERENCES

[1] Japan Student Services Organization, *Survey on the status of International Students in 2023*, Accessed: 2025-01-29, 2024. [Online]. Available: https://www.studyinjapan.go.jp/ja/statistics/enrollment/data/2405241100.html.

[2] Hinton, Geoffrey and Deng, Li and Yu, Dong and Dahl, George E and Mohamed, Abdel-rahman and Jaitly, Navdeep and Senior, Andrew and Vanhoucke, Vincent and Nguyen, Patrick and Sainath, Tara N and others, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, Łukasz and Polosukhin, Illia, "Attention Is All You Need," in *Proc. of Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[4] Kenton, Jacob Devlin Ming-Wei Chang and Toutanova, Lee Kristina, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[5] Y. Liu, "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692.

[6] Chowdhery, Aakanksha and Narang, Sharan and Devlin, Jacob and Bosma, Maarten and Mishra, Gaurav and Roberts, Adam and Barham, Paul and Chung, Hyung Won and Sutton, Charles and Gehrmann, Sebastian and others, "PaLM: Scaling Language Modeling with Pathways," *Journal of Machine Learning Research*, vol. 24, 240:1–240:113, 2023. [Online]. Available: https://jmlr.org/papers/v24/22-1144.html.

[7] Brown, Tom and Mann, Benjamin and Ryder, Nick and Subbiah, Melanie and Kaplan, Jared D and Dhariwal, Prafulla and Neelakantan, Arvind and Shyam, Pranav and Sastry, Girish and Askell, Amanda and others, "Language Models Are Few-Shot Learners," in *Proc. of Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[8] Chaoyou Fu and Peixian Chen and Yunhang Shen and Yulei Qin and Mengdan Zhang and Xu Lin and Jinrui Yang and Xiawu Zheng and Ke Li and Xing Sun and Yunsheng Wu and Rongrong Ji, "MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models," *arXiv*, vol. abs/2306.13394, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:259243928.

[9] Alexey Dosovitskiy and Lucas Beyer and Alexander Kolesnikov and Dirk Weissenborn and Xiaohua Zhai and Thomas Unterthiner and Mostafa Dehghani and Matthias Minderer and Georg Heigold and Sylvain Gelly and Jakob Uszkoreit and Neil Houlsby, "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. of the International Conference on Learning Representations*, 2021.

[10] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] Shi, Baoguang and Bai, Xiang and Yao, Cong, "An End-to-End Trainable Neural Network for Image-Based Sequence Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[12] Galal M. Binmakhashen and Sabri A. Mahmoud, "Document Layout Analysis: A Comprehensive Survey," *ACM Computing Surveys*, vol. 52, no. 6, pp. 1–36, 2019.

[13] Google, *Google Translate*, Accessed: 2025-02-20, 2025. [Online]. Available: https://translate.google.com/.

[14] Microsoft Support, *Present from PowerPoint Live in Microsoft Teams*, Accessed: 2025-02-20, 2025. [Online]. Available: https://support.microsoft.com/en-us/office/present-from-powerpoint-live-in-microsoft-teams-%5C%5C28b20e74-7165-499c-9bd4-0ad975d448ad.

[15] K. Sudo, T. Hayashi, Y. Nishimura, and S. Nakamura, "Prototype of an Automatic Translation Subtitle Generation System for Lecture Archives," in *IPSJ SIG Notes*, vol. 2019-NL-240, 2019, pp. 1–4.

[16] A. Bérard, L. Besacier, A. C. Kocabiyikoglu, and O. Pietquin, "End-to-End Automatic Speech Translation of Audiobooks," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6224–6228.

[17] Katharina Anderer and Andreas Reich and Matthias Wölfel, "MaViLS: A Benchmark Dataset for Video-to-Slide Alignment," in *Proc. of Interspeech 2024*, 2024, pp. 1375–1379.

[18] Hotta, Makoto and Leow, Chee Siang and Kitaoka, Norihide and Nishizaki, Hiromitsu, "Evaluation of Speech Translation Subtitles Generated by ASR with Unnecessary Word Detection," in *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, 2024, pp. 815–819.

[19] OpenAI, *Introducing Whisper*, This reference is for Whisper Large v3. Accessed: 2025-02-01, 2022. [Online]. Available: https://openai.com/index/whisper/.

[20] DeepL, *DeepL Translator*, Accessed: 2025-01-29, 2025. [Online]. Available: https://www.deepl.com/en/translator.

[21] Nils Reimers and Iryna Gurevych, *SentenceTransformers Documentation*, Accessed: 2025-02-01, 2025. [Online]. Available: https://sbert.net/.

[22] The Japan Foundation and Japan Educational Exchanges and Services. "Japanese-Language Proficiency Test (JLPT) Official Website." (2025), [Online]. Available: https://www.jlpt.jp/ (visited on 07/09/2025).

[23] OpenAI, *ChatGPT API Documentation*, Accessed: 2025-02-06, 2025. [Online]. Available: https://platform.openai.com/docs/.

[24] DeepL GmbH, *DeepL API Documentation*, Accessed: 2025-02-06, 2025. [Online]. Available: https://www.deepl.com/docs-api.

[25] OpenCV, *Open Source Computer Vision Library*, Accessed: 2025-02-20, 2025. [Online]. Available: https://opencv.org/.

[26] Kotaro Kinoshita, *YomiToku: AI-Powered Japanese Document Analysis*, Accessed: 2025-02-01, 2025. [Online]. Available: https://kotaro-kinoshita.github.io/yomitoku/.

[27] Beijing Academy of Artificial Intelligence, *BGE-M3: Multi-Functional, Multi-Lingual, and Multi-Granular Embedding Model*, Accessed: 2025-02-01, 2024. [Online]. Available: https://huggingface.co/BAAI/bge-m3.

[28] Sashi Novitasari and Sakriani Sakti and Satoshi Nakamura, "Neural Incremental Speech Recognition Toward Real-Time Machine Speech Translation," *IEICE Transactions on Information and Systems*, vol. E104.D, no. 12, pp. 2195–2208, 2021.