BOOM: *Beyond Only One Modality*
KIT's Multimodal Multilingual Lecture Companion

Sai Koneru, Fabian Retkowski, Christian Huber, Lukas Hilgert,
Seymanur Akti, Enes Yavuz Ugan, Alexander Waibel, Jan Niehues
Karlsruhe Institute of Technology

firstname.lastname@kit.edu

Abstract

The globalization of education and rapid growth of online learning have made localizing educational content a critical challenge. Lecture materials are inherently multimodal, combining spoken audio with visual slides, which requires systems capable of processing multiple input modalities. To provide an accessible and complete learning experience, translations must preserve all modalities: text for reading, slides for visual understanding, and speech for auditory learning. We present **BOOM**, a multimodal multilingual lecture companion that jointly translates lecture audio and slides to produce synchronized outputs across three modalities: translated text, localized slides with preserved visual elements, and synthesized speech. This end-to-end approach enables students to access lectures in their native language while aiming to preserve the original content in its entirety. Our experiments demonstrate that slide-aware transcripts also yield cascading benefits for downstream tasks such as summarization and question answering. The demo video and code can be found at https://ai4lt.github.io/boom/[1].

1 Introduction

Access to educational content in a learner's native language greatly enhances the learning experience for university students. Localizing lecture material reduces communication barriers, improves accessibility, and enables learners to engage more deeply with complex concepts. As higher education becomes increasingly global, the ability to provide multilingual lecture content both in-person and online has become essential to increase accessibility to educational resources (Muthuswamy and Varshika, 2023; Gambier, 2023).

With the ongoing digitalization of teaching, lecture content itself is inherently multimodal. The primary modality is the lecture audio, which can be converted into transcripts via Automatic Speech Recognition (ASR) (Pham et al., 2019; Radford et al., 2022). Instructional material is presented through slides, and additional outputs, such as summaries, chapters, and question–answer interactions, can be generated based on the transcript in a cascaded setup using modern Large Language Model (LLM)-based systems to enhance the learning experience (Waibel and Fuegen, 2012; Waibel, 2014; Anderer et al., 2025; Retkowski et al., 2025). To ensure accessibility for all students, including non-native speakers, these outputs should also be available in multiple languages. Effective localization must therefore handle this diversity of content, spanning audio, text, and visual materials, making lecture translation a truly multimodal challenge.

This multimodality introduces complexity but also offers valuable contextual signals. Images often contain additional cues ranging from scene information in natural images and definitions, formulas, diagrams, and domain-specific terminology in slides that help disambiguate spoken content (Nguyen et al., 2025) and support downstream tasks such as Summarization (SUM) and Question Answering (QA). Leveraging these visual cues enables translation systems to move beyond audio-only processing and incorporate richer semantic information throughout the lecture translation pipeline (Waibel, 2018; Chen et al., 2024; Sinhamahapatra and Niehues, 2025).

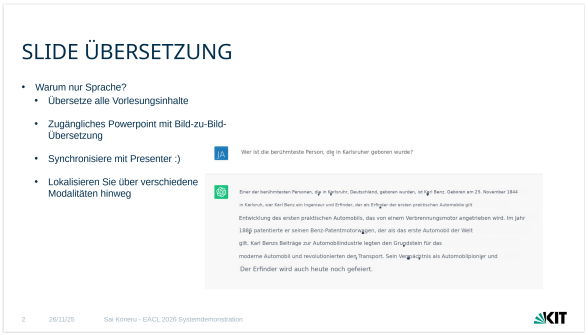
Machine Translation (MT) forms the foundation of localization, evolving from rule-based systems (Hutchins, 2004) to Neural MT (NMT; Vaswani et al. 2017; Koehn and Knowles 2017; Johnson et al. 2017) and then Speech Translation (ST), which directly translates spoken content. Modern ST handles many languages (Barrault et al., 2023) but often processes short segments, limiting context and potential to benefit from multimodality.

In this work, we address multimodality on both

[1]All released code and models are licensed under the MIT License

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 3: System Demonstrations, pages 175–187

(a) Original English Slide



(b) Translated German Slide

Figure 1: Comparison of the English (original) and German (translated) slides. Text outside the images is translated with a unimodal system for efficiency, while text inside the images is translated using a multimodal system.

the input and output sides of lecture localization. On the input side, we incorporate slide screenshots into the ST pipeline to provide contextual grounding that improves translation accuracy and downstream LLM performance. On the output side, we tackle the challenge of localizing lecture slides themselves. Slides often contain text embedded within images, such as diagram labels, equations, or annotations, that existing ST tools typically ignore. Localizing such material requires detecting, recognizing, translating, and re-rendering text while preserving layout, alignment, font style, and visual coherence (illustrated in Figure 1).

To overcome these limitations, we extend the Lecture Translator (LT) software (Huber et al., 2023) with OmniFusion (Koneru et al., 2025), a multilingual multimodal ST model that uses slide images to enrich translation. We further introduce a fully open-source slide translation system capable of translating text inside slide images and rendering it back into its original layout, enabling complete slide localization. Together, these components form a unified multimodal lecture localization pipeline that combines improved ST with synchronized slide translation, significantly enhancing accessibility for learners across languages.

Our main contributions include:

- Adapt and integrate OmniFusion to leverage lecture slide screenshots during live translation, by extracting relevant slides from segmented audio.
- Introduce an open-source image-to-image translation pipeline with modular components, enabling future research on full-image/slide translation and rendering.
- Demonstrate the impact of including images on ST for downstream NLP tasks across different LLMs, showing performance improvements in different language pairs. We also evaluate several optical character recognition (OCR) models and the translation quality of unimodal and multimodal NMT models for image translation.

2 System Description

To fully localize lecture content, including audio and slides, across multiple modalities and languages, and to support accessibility tasks such as SUM and QA, we develop multimodal translation systems. Our approach performs multimodal ST and leverages the resulting transcripts for downstream LLM tasks. We also translate slides by converting text and images into the target language while preserving their layout and visual coherence.

To map visual context to each audio segment and improve usability, we built a PDF viewer that displays slides with overlaid captions synchronized to the presenter's selected slide (Figure 8). This interface enables participants to follow translations while viewing slides and allows the system to automatically identify which slide corresponds to each audio segment, providing essential context for multimodal ST.

In this section, we first describe the multimodal ST pipeline, including how slide images are extracted and associated with audio segments. Next, we outline how the resulting translations are used for downstream SUM and QA. Finally, we present the slide-translation process, detailing how text embedded within slide images is detected, translated, and re-rendered. Additionally, details about Text-to-Speech (TTS) are described in Appendix A.1.

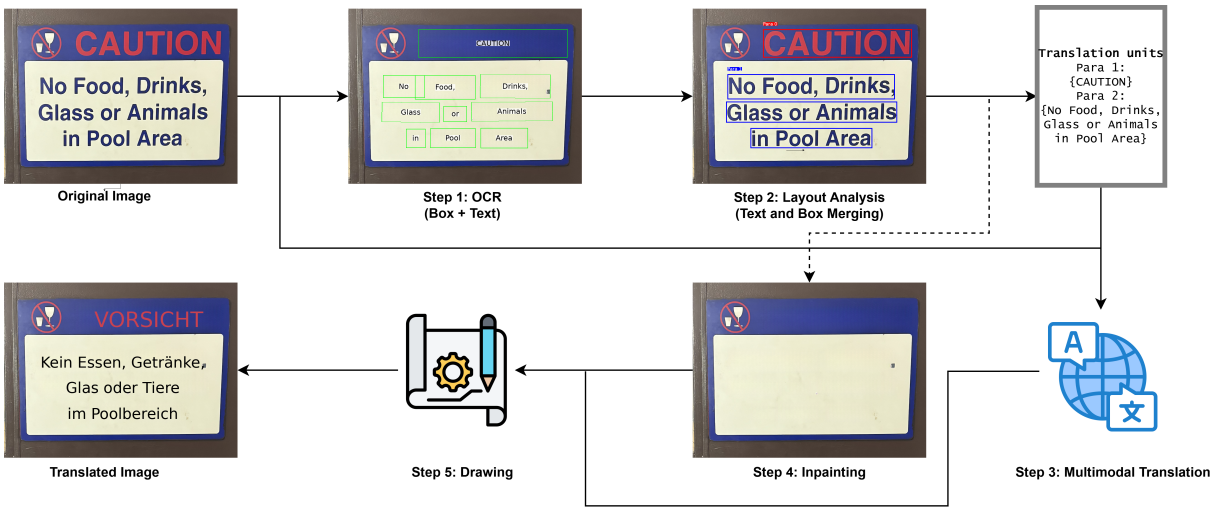


Figure 2: Overview of the image translator pipeline. Arrows indicate the inputs to each step. All steps are model-based except for drawing, which uses heuristic rules.

2.1 Multimodal Speech Translation

Several ST systems support translation across multiple languages, but they are not directly suitable for live lectures. Most are trained for offline tasks with fixed segmentation, which is incompatible with streaming audio, and require simultaneous translation policies (Niehues et al., 2016, 2018; Polák et al., 2023) to determine when enough audio has been received. Existing systems are either unimodal, ignoring slides, or multimodal but lack multilingual support. Lecture scenarios demand both multimodality and multilinguality.

To address these challenges, we adopt the OmniFusion model for multimodal ST, which supports multiple languages and has been shown to improve quality when integrating slides. Since it is trained primarily on clean speech, we fine-tune it on noisy data[2]. For streaming translation, we follow the LT policy (Huber et al., 2023), combining voice-activity detection with Local-Agreement to produce low-latency outputs.

Accurate visual context is crucial for effective translation. The PDF viewer tracks the slide displayed during each audio segment, allowing us to extract a screenshot from the middle of the segment and feed it to the ST model. This provides relevant visual cues, improving translation quality, especially for technical content, while enabling participants to follow translations in real time.

[2] https://huggingface.co/skoneru/OmniFusion_v2

2.1.1 Summarization & Question-Answering

Beyond translating spoken content, lecture material should also be chaptered (Zechner and Waibel, 2000a,b; Schneider et al., 2025; Retkowski and Waibel, 2024), meaning split into coherent functional and semantic sections, and then summarized in multiple languages and made available for interactive QA. To support these tasks, we use the transcribed multimodal ST output as context. Although modern LLMs can handle long contexts efficiently, their context window is still limited, so we adopt the following strategy.

For summarization, lectures are first translated into multiple languages. Each lecture is then divided into chapters, which prevents context-window overflow and also produces conceptually cleaner summaries, since chapters contain locally coherent content and avoid the topic drift that often appears in global summaries. For each chapter, we generate several forms of compressed representations. These include transcript compressions at multiple ratios such as 50 percent, 70 percent, and 90 percent, as well as length-controlled summaries whose size is determined by the length of the source section (Retkowski and Waibel, 2025). All summaries are first produced in English to benefit from the stronger performance of LLMs on English text and are then translated into the target languages.

For QA, we follow a similar approach: the English transcript, organized by chapters, is used with Retrieval-Augmented Generation (RAG) to query an LLM (Anderer et al., 2025), and the resulting answers are translated into the target languages.

Model	CER (↓)	TER (↓)	Sub.	Del.	Ins.	Average Time (↓) (Seconds)
EasyOCR	56.44	57.44	1488	29337	553	0.22
Paddle-OCR-v4	11.31	16.53	880	2791	2435	0.06
Paddle-OCR-v5	13.48	16.91	1717	2639	3014	0.10
Qwen-2.5-VL 7B	13.54	12.77	413	2348	3144	5.10

Table 1: Performance of OCR models on the VISTRA benchmark. Evaluations are restricted to English text in signboards and similar visual contexts, and therefore do not reflect performance across broader OCR domains.

2.2 Slide Translation

Another challenge for making lectures accessible is translating slides into multiple languages. Slides contain both editable text and images with embedded text. For editable text, we use a Python-based PowerPoint parser[3] to extract text blocks and translate them with standard unimodal MT, avoiding multimodal models due to computational cost.

Text inside images cannot be directly extracted, often lacks surrounding linguistic context, and relies on visual elements for interpretation, making multimodal translation necessary. After translation, text must be reinserted into the original image to preserve layout and visual meaning. To address this, we propose an **image-translation pipeline** that detects, recognizes, translates, and re-renders text within slide images (Figure 2).

2.2.1 Optical Character Recognition

The system begins with extracting text from slide images using PaddleOCR v5 (Cui et al., 2025), which supports multiple languages and outputs both recognized text and bounding boxes, typically at the word or character level. While sufficient for translation, these detections do not form coherent segments or preserve semantic structure, requiring layout analysis.

2.2.2 Layout Analysis

We then apply layout analysis using the Hi-SAM model[4] (Ye et al., 2025b), which predicts block-level regions and their constituent lines. OCR boxes are grouped into block-level and line-level segments, producing sentence-like units suitable for translation. Layout analysis also preserves structural cues, such as grouping, font size, and color, that aid re-rendering. For instance, bullet list items or diagram labels are grouped to maintain consistent formatting.

2.2.3 Multimodal Translation

Text from each block is concatenated and translated using OmniFusion adapted from Qwen Omni 2.5 7B(Ye et al., 2025a) and SeedX PPO 7B (Cheng et al., 2025), which leverages the slide image as visual context. This multimodal approach is particularly helpful for short, ambiguous, or visually grounded text.

2.2.4 Inpainting

Before inserting the translated text, the original text regions are removed using Simple-LaMa[5] (Suvorov et al., 2021), a lightweight inpainting model that reconstructs the background with minimal artifacts, preserving slide quality.

2.2.5 Drawing

Translated text is then rendered back onto the slide. Fully automatic diffusion-based methods proved unsuitable because repeated edits gradually degraded clarity. Instead, a heuristic drawing module estimates original text styling and positions the translated text within the same layout and line structure. This preserves alignment, spatial organization, and overall visual coherence, ensuring the localized slide matches the structure and intent of the original.

3 Experiments

3.1 Evaluation Data & Metrics

Since no dataset directly provides lecture slides with ground-truth translations, summaries, and QA pairs, we evaluate our approach on established benchmarks that approximate these tasks. For image translation, we use the VISTRA benchmark (Salesky et al., 2024), which contains real-world images such as street signs with ground-truth OCR and translations for English → German, Chinese,

[3] https://pypi.org/project/python-pptx/
[4] *sam_vit_l_0b3195.pth*
[5] https://github.com/enesmsahin/simple-lama-inpainting/

Model	de			es			ru			zh		
	BLEU (↑)	ChrF (↑)	COMET (↑)	BLEU (↑)	ChrF (↑)	COMET (↑)	BLEU (↑)	ChrF (↑)	COMET (↑)	BLEU (↑)	ChrF (↑)	COMET (↑)
OCR Predicted + Line-level												
SeedX 7B PPO	6.7	21.3	50.9	18.3	48.8	68.9	10.8	37.8*	65.6*	0.6	7.4	62.8
Tower-Instruct 7B	4.5	23.3	50.5	11.6	40.0	63.2	7.3	28.9	59.1	3.5*	17.2	63.1
OmniFusion	9.2*	25.3*	53.5*	19.8*	50.7*	70.4*	11.0*	34.8	64.6	1.3	22.1*	67.6*
OCR Predicted + Layout-level												
SeedX 7B PPO	10.3	23.7	53.1	28.4*	56.8*	74.0	17.3*	43.4*	71.2*	2.0	14.8	67.8
Tower-Instruct 7B	11.2	27.4	53.7	19.1	46.4	68.2	10.6	30.7	63.3	8.4*	22.9	68.3
OmniFusion	13.6*	30.1*	56.9*	28.1	56.2	74.5*	15.2	36.7	68.5	5.4	27.9*	71.4*
Ground-Truth (OCR + Segmentation)												
SeedX 7B PPO	14.5	27.4	57.8	35.6	**63.1***	81.8	**23.5***	**49.2***	**78.9***	13.9	34.5	83.4
Tower-Instruct 7B	11.0	31.6	59.2	28.1	53.2	75.4	15.1	34.4	69.2	**23.3***	37.5	83.1
OmniFusion	**18.4***	**35.0***	**62.2***	**36.9***	62.5	**81.9***	20.4	38.8	74.0	16.5	**43.5***	**84.6***

Table 2: Comparison of translation quality across models on the VISTRA benchmark. OCR-predicted results rely on PaddleOCR-v5. The best score within each evaluation setting is marked with *, and the best overall is **bold**.

Russian, Spanish. OCR performance is measured using Character Error Rate (CER), Term Error Rate (TER; Snover et al. 2006), and latency. Translation quality is evaluated with BLEU, ChrF using SacreBLEU[6] (Post, 2018), and COMET[7] (Rei et al., 2022). For downstream tasks, we use the MCIF dataset of ACL talks (Papi et al., 2025b) and report normalized BERTScore to evaluate generated summaries and answers.

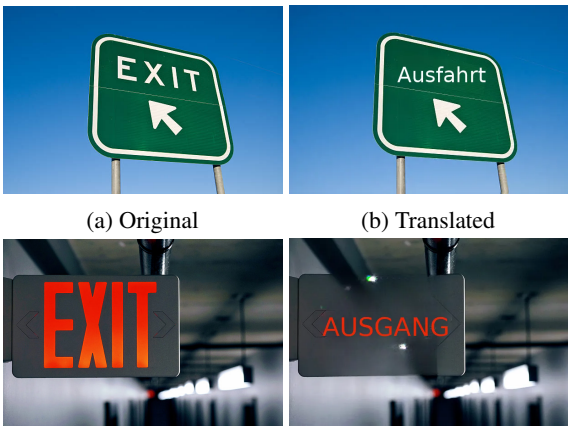
3.2 Image Translation

We evaluate our complete image-translation pipeline along three dimensions: OCR accuracy, translation quality, and component runtime.

OCR Evaluation. Table 1 summarizes OCR performance of several open-source systems and the vision LLM Qwen-2.5-VL (7B; Bai et al. 2025). EasyOCR[8] performs the worst due to its lightweight and less robust design. PaddleOCR v4 and v5 achieve similar and much higher accuracy, while Qwen-2.5-VL matches PaddleOCR but suffers from very high latency (0.1s → 5s per image). Considering accuracy, latency, and language coverage, PaddleOCR v5 provides the best trade-off and is used for all subsequent experiments.

Translation Quality Table 2 presents translation results for both unimodal LLMs, Tower 7B (Alves et al., 2024) and SeedX, and the multimodal OmniFusion model. To evaluate the impact of input segmentation, we compare line-level segmentation (where each OCR line is treated independently), block-level segmentation (where lines are grouped within layout regions), and ground-truth OCR plus segmentation as an upper bound.

Overall, OmniFusion consistently outperforms unimodal translation in most languages, showing that visual context from images helps disambiguate short or visually grounded text, such as diagram labels or signs. Ground-truth OCR and segmentation yield the best performance, highlighting the importance of accurate text extraction and layout grouping. Block-level segmentation improves translation over line-level segmentation, confirming that coherent sentence-like units are critical for high-quality output. Unimodal translation performs better in Russian, indicating potentially less reliance on visual context in this direction.



(a) Original (b) Translated

(c) Original (d) Translated

Figure 3: Example illustrating that our Image Translator uses context for disambiguation. The word "Exit" can mean "Ausgang" in the context of a pedestrian exit and "Ausfahrt" in the context of a car exit. Our translator correctly leverages the visual context to produce different translations, even when the source text is identical in both scenarios.

Table 4 in Appendix A shows inference times for different components. Layout analysis and translation are slowest, whereas OCR and image ren-

[6] nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.3.1
[7] Unbabel/wmt22-comet-da
[8] https://github.com/JaidedAI/EasyOCR

Language	ST Input	LLaMA 3.1 8B	GPT OSS 20B	Mistral Small 3.2 24B
Summarization				
English	🔊	18.4	12.1	18.1
	🔊🖼	20.5	12.7	19.7
German	🔊	20.6	18.0	21.7
	🔊🖼	23.4	18.9	24.1
Italian	🔊	22.5	18.9	25.4
	🔊🖼	24.4	19.7	26.3
Chinese	🔊	35.7	31.9	35.9
	🔊🖼	35.3	31.7	35.8
Question Answering				
English	🔊	31.5	23.0	34.5
	🔊🖼	34.5	22.0	35.4
German	🔊	32.0	21.5	37.2
	🔊🖼	33.6	22.5	37.6
Italian	🔊	33.7	19.4	36.2
	🔊🖼	34.7	20.5	34.7
Chinese	🔊	35.8	30.5	32.4
	🔊🖼	35.4	30.0	32.7

Table 3: Summarization and Question Answering performance of different LLMs on the MCIF test dataset based on translations of the presentations with OmniFusion. Reported is BERTScore (↑), rescaled with the baseline. 🔊: Audio only, 🔊🖼: Audio + Image.

dering add relatively minor overhead, suggesting that optimizing efficiency for these would provide the largest latency gains. Figure 3 illustrates an example in which multimodal translation disambiguates text using visual context, demonstrating the practical benefit of incorporating images.

3.3 Downstream Tasks

We analyze how downstream performance on the MCIF benchmark (Papi et al., 2025b), specifically for Summarization and Question Answering, is affected when the transcript used as context is generated by the multimodal speech-translation system. Using the task instructions provided by MCIF, we prompt each evaluated model directly with the translated talk transcript produced by our pipeline. We evaluate three LLMs: LLaMA 3.1 8B (Grattafiori et al., 2024), GPT-OSS 20B (OpenAI, 2025), and Mistral-Small 3.2 24B (Jiang et al., 2023) [9] [10]. This setup allows us to measure how using audio-only transcripts compared to multimodal transcripts that also incorporate slide information influences downstream task performance.

Summarization. As shown in Table 3, summaries generated from audio+image input (🔊🖼) consistently outperform those based on audio-only (🔊) across most languages and models, even though the summarization models are text-only. The gains are most pronounced in English, German and Italian, while results for Chinese slightly degraded. We presume this is because English domain terminology appears in references for Latin-alphabet languages, while the lexical distance between English and Chinese prevents the models from consistently benefiting from additional context provided in English language.

Question Answering. In most settings, the results for QA are slightly better when incorporating visual context, though the gains are much less pronounced compared to summarization. In most cases, we observe small gains but also performance regression in four out of twelve language–model combinations. We assume that we do not see higher improvements and regression because the LLM does not receive the image data itself but just the (through multimodality improved) textual context which is not enough for the model to answer the questions more reliably.

4 Related Work

Streaming ST has been extensively studied in the last decade (Macháček et al., 2023; Guo et al., 2025; Papi et al., 2025a). Several lecture translation tools have also leveraged ST (Cho et al., 2013; Niehues et al., 2016; Son Nguyen et al., 2020; Müller et al., 2016; Dessloch et al., 2018; Huber et al., 2023), but these systems primarily rely on audio input. In contrast, our work extends lecture translation to multimodal input, incorporating visual cues from slides, and multimodal output, producing translated audio and slides in multiple languages.

Image-to-image translation remains relatively under-explored. Several research works focus on road sign translation (Gao et al., 2001; Yang et al., 2001; Zhang et al., 2002; Chen et al., 2002, 2004) facing many similar challenges to translating images in academic slies. Interest in this area is growing with the availability of larger datasets (Zuo et al., 2025; Li et al., 2025; Zhuang et al., 2025), but most existing work focuses solely on text translation within images, without addressing the aligned re-rendering of the visual content. An initial step in this direction is (Tian et al., 2025), which explicitly models the rendering process. Our image translation pipeline provides a modular foundation, enabling researchers to integrate models at any stage from OCR to translation and rendering, without

[9] https://mistral.ai/news/mistral-small-3-1

[10] https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506

needing to implement additional components.

5 Conclusion

This paper presents a multimodal, multilingual lecture translation system that leverages multiple input modalities to generate translations across different output modalities. Future work includes conducting human evaluations to assess the quality of translated slides and audio, enabling targeted improvements to the system.

Limitations

To assess the effectiveness of our slide translation, we use the VISTRA benchmark as a proxy. However, this benchmark does not fully reflect translation quality in the lecture domain, nor does it allow us to evaluate the quality of rendered slides. Human evaluation is therefore needed to assess the rendering quality of translated slides, including layout preservation and visual coherence. For SUM and QA, we conduct evaluation only after the entire talk has been translated, which does not accurately simulate a live lecture scenario. Benchmarks with questions aligned to the lecture timeline would provide more realistic and informative evaluations for our use-case.

Acknowledgments

The research leading to these results was supported by European Union's Horizon Europe programme grant agreement No. 101213369 (DVPS) and No. 101135798 (Meetween), The German Federal Ministry of Education, Research (BMBF) under the Robotics Institute Germany (RIG) and "How is AI Changing Science? Research in the Era of Learning Algorithms" (HiAICS) project.

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G C de Souza, and André F T Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv [cs.CL]*.

Katharina Anderer, Karin Müller, Lukas Strobel, Matthias Wölfel, Jan Niehues, and Kathrin Gerling. 2025. Making lecture videos accessible for students who are blind or have low vision through ai-assisted navigation and visual question answering. In *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '25, New York, NY, USA. Association for Computing Machinery.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, and 1 others. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Xilin Chen, Jie Yang, Jing Zhang, and Alex Waibel. 2002. Automatic detection of signs with affine transformation. In *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, pages 32–36. IEEE.

Xilin Chen, Jie Yang, Jing Zhang, and Alex Waibel. 2004. Automatic detection and recognition of signs from natural scenes. *IEEE Transactions on image processing*, 13(1):87–99.

Zhe Chen, Heyang Liu, Wenyi Yu, Guangzhi Sun, Hongcheng Liu, Ji Wu, Chao Zhang, Yu Wang, and Yanfeng Wang. 2024. M^{3}AV: A multimodal, multigenre, and multipurpose audio-visual academic lecture dataset. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060, Bangkok, Thailand. Association for Computational Linguistics.

Shanbo Cheng, Yu Bao, Qian Cao, Luyang Huang, Liyan Kang, Zhicheng Liu, Yu Lu, Wenhao Zhu, Jingwen Chen, Zhichao Huang, Tao Li, Yifu Li, Huiying Lin, Sitong Liu, Ningxin Peng, Shuaijie She, Lu Xu, Nuo Xu, Sen Yang, and 7 others. 2025. Seed-X: Building strong multilingual translation LLM with 7B parameters. *arXiv [cs.CL]*.

Eunah Cho, Christian Fügen, Teresa Herrmann, Kevin Kilgour, Mohammed Mediani, Christian Mohr, Jan Niehues, Kay Rottmann, Christian Saam, Sebastian Stüker, and 1 others. 2013. A real-world system for simultaneous translation of german lectures. In *INTERSPEECH*, pages 3473–3477.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. Paddleocr 3.0 technical report. *Preprint*, arXiv:2507.05595.

Florian Dessloch, Thanh-Le Ha, Markus Müller, Jan Niehues, Thai-Son Nguyen, Ngoc-Quan Pham, Elizabeth Salesky, Matthias Sperber, Sebastian Stüker, Thomas Zenkel, and 1 others. 2018. Kit lecture translator: Multilingual speech translation with one-shot learning. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 89–93.

Yves Gambier. 2023. Audiovisual translation and multimodality: What future? *Media and intercultural communication: a multidisciplinary journal.*, 1(1):1–16.

Jiang Gao, Jie Yang, Ying Zhang, and Alex Waibel. 2001. Text detection and translation from natural scenes. Technical report.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv [cs.AI]*.

Shoutao Guo, Xiang Li, Mengge Liu, Wei Chen, and Yang Feng. 2025. Streamuni: Achieving streaming speech translation with a unified large speech-language model. *arXiv preprint arXiv:2507.07803*.

Christian Huber, Tu Anh Dinh, Carlos Mullov, Ngoc-Quan Pham, Thai-Binh Nguyen, Fabian Retkowski, Stefan Constantin, Enes Ugan, Danni Liu, Zhaolin Li, and 1 others. 2023. End-to-end evaluation for low-latency simultaneous speech translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 12–20.

W John Hutchins. 2004. The georgetown-ibm experiment demonstrated in january 1954. In *Conference of the Association for Machine Translation in the Americas*, pages 102–114. Springer.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Melvin Johnson, Mike Schuster, Quoc Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, and 1 others. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*. Association for Computational Linguistics.

Sai Koneru, Matthias Huck, and Jan Niehues. 2025. Omnifusion: Simultaneous multilingual multimodal translations via modular fusion. *Preprint*, arXiv:2512.00234.

Jungil Kong, Jihoon Park, Beomjeong Kim, Jeongmin Kim, Dohee Kong, and Sangjin Kim. 2023. Vits2: Improving quality and efficiency of single-stage text-to-speech with adversarial learning and architecture design. *arXiv preprint arXiv:2307.16430*.

Bo Li, Shaolin Zhu, and Lijie Wen. 2025. MIT-10M: A large scale parallel corpus of multilingual image translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5154–5167, Abu Dhabi, UAE. Association for Computational Linguistics.

Dominik Macháček, Raj Dabre, and Ondřej Bojar. 2023. Turning whisper into real-time transcription system. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics (ACL).

Markus Müller, Sarah Fünfer, Sebastian Stüker, and Alex Waibel. 2016. Evaluation of the kit lecture translation system. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1856–1861.

Vimala Venugopal Muthuswamy and G Varshika. 2023. Analysing the influence of cultural distance and language barriers on academic performance among international students in higher education institutions. *Journal of International Students*, 13(3):415–440.

Thai-Binh Nguyen, Ngoc-Quan Pham, and Alexander Waibel. 2025. Cocktail-party audio-visual speech recognition. In *Proc. Interspeech 2025*, pages 1828–1832.

Jan Niehues, Thai Son Nguyen, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Müller, Matthias Sperber, Sebastian Stüker, and Alex Waibel. 2016. Dynamic transcription for low-latency speech translation. In *Interspeech*, pages 2513–2517.

Jan Niehues, Ngoc-Quan Pham, Thanh-Le Ha, Matthias Sperber, and Alex Waibel. 2018. Low-latency neural speech translation. *arXiv preprint arXiv:1808.00491*.

OpenAI. 2025. gpt-oss-120b & gpt-oss-20b model card. *CoRR*, abs/2508.10925.

Sara Papi, Peter Polák, Dominik Macháček, and Ondřej Bojar. 2025a. How "real" is your real-time simultaneous speech-to-text translation system? *Trans. Assoc. Comput. Linguist.*, 13:281–313.

Sara Papi, Maike Züfle, Marco Gaido, Beatrice Savoldi, Danni Liu, Ioannis Douros, Luisa Bentivogli, and Jan Niehues. 2025b. Mcif: Multimodal crosslingual instruction-following benchmark from scientific talks. *Preprint*, arXiv:2507.19634.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, Sebastian Stüker, and Alexander Waibel. 2019. Very deep self-attention networks for end-to-end speech recognition. *arXiv preprint arXiv:1904.13377*.

Peter Polák, Brian Yan, Shinji Watanabe, Alex Waibel, and Ondřej Bojar. 2023. Incremental blockwise beam search for simultaneous speech translation with controllable quality-latency tradeoff. *arXiv preprint arXiv:2309.11379*.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fabian Retkowski and Alexander Waibel. 2024. From text segmentation to smart chaptering: A novel benchmark for structuring video transcriptions. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 406–419, St. Julian's, Malta. Association for Computational Linguistics.

Fabian Retkowski and Alexander Waibel. 2025. Zero-shot strategies for length-controllable summarization. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 551–572, Albuquerque, New Mexico. Association for Computational Linguistics.

Fabian Retkowski, Maike Züfle, Andreas Sudmann, Dinah Pfau, Shinji Watanabe, Jan Niehues, and Alexander Waibel. 2025. Summarizing speech: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27263–27294, Suzhou, China. Association for Computational Linguistics.

Elizabeth Salesky, Philipp Koehn, and Matt Post. 2024. Benchmarking visually-situated translation of text in natural images. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1167–1182.

Felix Schneider, Marco Turchi, and Alex Waibel. 2025. Policies and evaluation for online meeting summarization. *arXiv preprint arXiv:2502.03111*.

Supriti Sinhamahapatra and Jan Niehues. 2025. Do slides help? multi-modal context for automatic transcription of conference talks. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16111–16121, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Thai Son Nguyen, Jan Niehues, Eunah Cho, Thanh-Le Ha, Kevin Kilgour, Markus Muller, Matthias Sperber, Sebastian Stueker, and Alex Waibel. 2020. Low latency asr for simultaneous speech translation. *arXiv e-prints*, pages arXiv–2003.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2021. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*.

Yanzhi Tian, Zeming Liu, Zhengyang Liu, Chong Feng, Xin Li, He-Yan Huang, and Yuhang Guo. 2025. Prim: Towards practical in-image multilingual machine translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13693–13708.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alexander Waibel. 2014. Translation and integration of presentation materials in cross-lingual lecture support. US Patent App. 14/302,149.

Alexander Waibel. 2018. Translation training with cross-lingual multi-media support. US Patent 9,892,115.

Alexander Waibel and Christian Fuegen. 2012. Simultaneous translation of open domain lectures and speeches. US Patent 8,090,570.

Jie Yang, Jiang Gao, Ying Zhang, Xilin Chen, and Alex Waibel. 2001. An automatic sign recognition and translation system. In *Proceedings of the 2001 workshop on Perceptive user interfaces*, pages 1–8.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, and 1 others. 2025a. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*.

Maoyuan Ye, Jing Zhang, Juhua Liu, Chenyu Liu, Baocai Yin, Cong Liu, Bo Du, and Dacheng Tao. 2025b.

Hi-sam: Marrying segment anything model for hierarchical text segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(03):1431–1447.

Klaus Zechner and Alex Waibel. 2000a. Diasumm: Flexible summarization of spontaneous dialogues in unrestricted domains. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*.

Klaus Zechner and Alex Waibel. 2000b. Minimizing word error rate in textual summaries of spoken language. In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.

Ying Zhang, Bing Zhao, Jie Yang, and Alex Waibel. 2002. Automatic sign translation. In *INTERSPEECH*, pages 645–648.

Wanru Zhuang, Wenbo Li, Zhibin Lan, Xu Han, Peng Li, and Jinsong Su. 2025. Patimt-bench: A multi-scenario benchmark for position-aware text image machine translation in large vision-language models. *arXiv preprint arXiv:2509.12278*.

Fei Zuo, Kehai Chen, Yu Zhang, Zhengshan Xue, and Min Zhang. 2025. InImageTrans: Multimodal LLM-based text image machine translation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20256–20277, Vienna, Austria. Association for Computational Linguistics.

A Appendix

Step	Time (seconds)
OCR	0.46
Layout Analysis	2.93
Multimodal Translation	3.10
Inpainting	0.42
Drawing	0.18

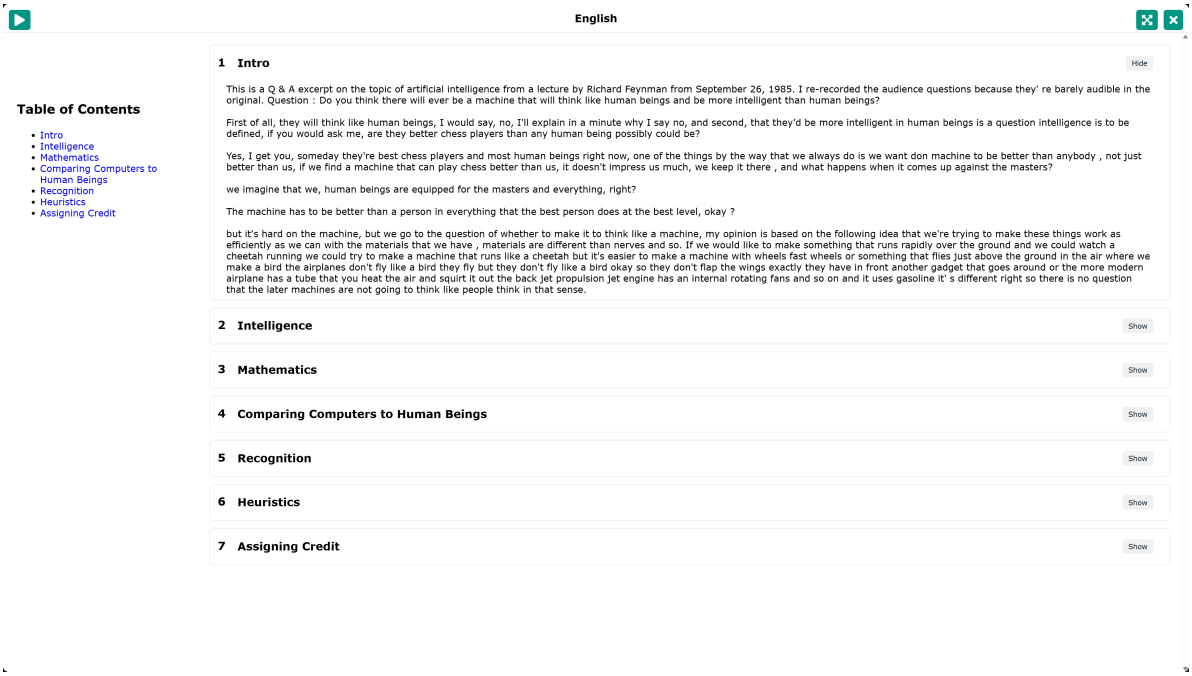
Table 4: Inference time for each step in the pipeline for translating the image shown in Figure 2.

A.1 Text-to-Speech

In Figure 5, the interface of the TTS output can be seen. It is possible to select between the simultaneous and consecutive modes. The simultaneous mode can be used when listening to the TTS output via headphones during the talk. The consecutive mode is suitable in dialog scenarios where the TTS output is paused as long as the system recognizes speaker input. We use the VITS/VITS2 (Kim et al., 2021; Kong et al., 2023) and Kokoro-82M to generate audio together with a rule-based streaming algorithm to segment input text into segments.

A.2 User Interface Screenshots

(a) English translation with segmentation into multiple chapters.



(b) German translation with segmentation into multiple chapters.

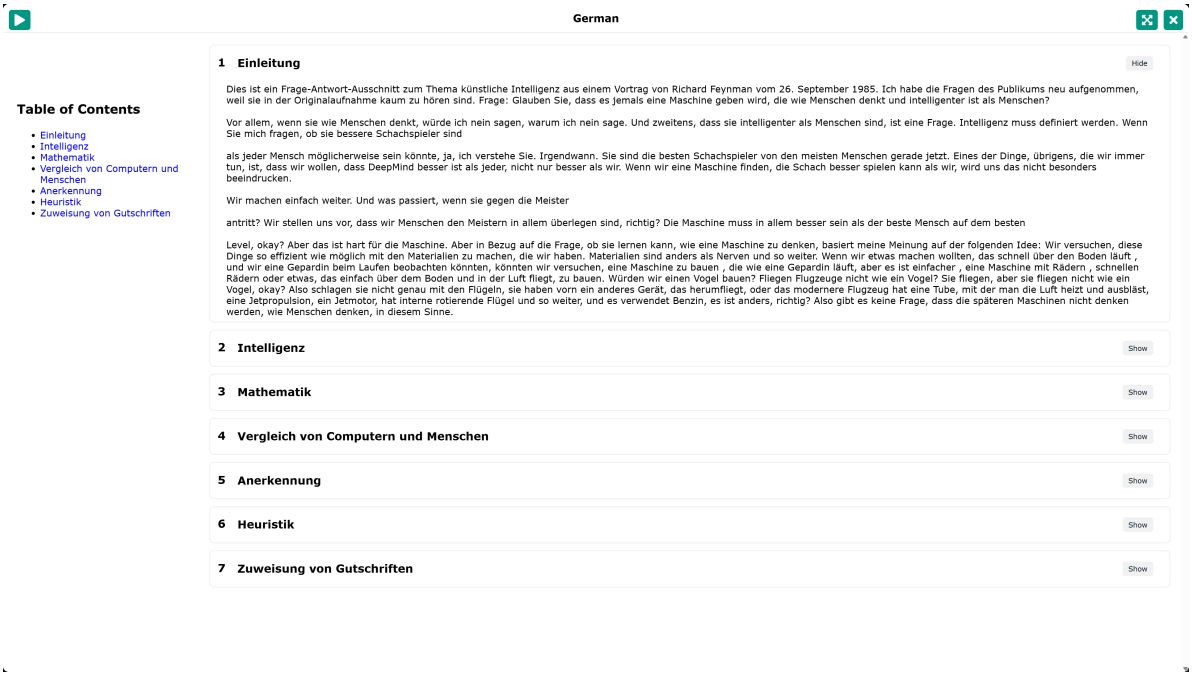


Figure 4: Translations of the YouTube video **"Richard Feynman: Can Machines Think?"** (https://www.youtube.com/watch?v=ipRvjS7q1DI). Subfigure (a) shows the English version; subfigure (b) shows the German version.

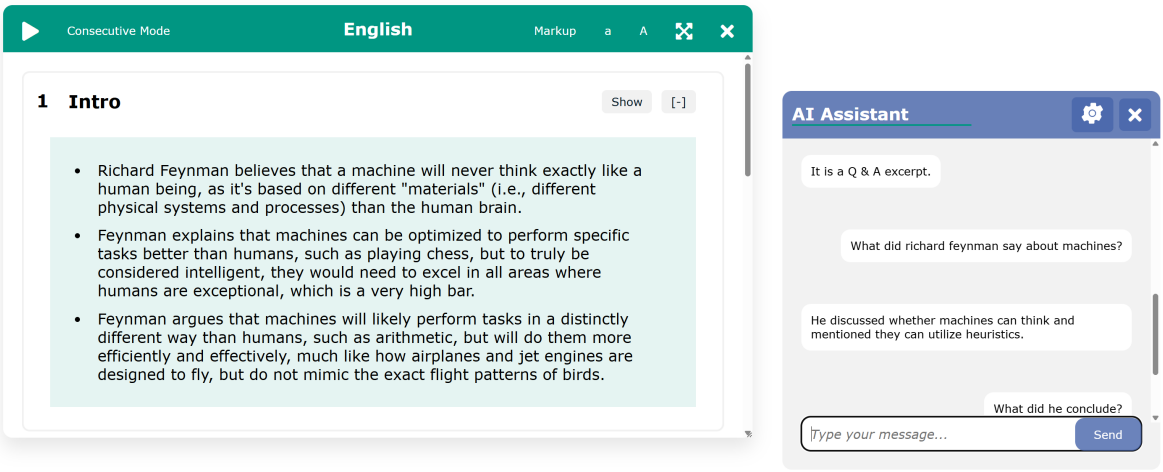


Figure 5: Summarization and Question Answering user interface. The summaries are shown for each chapter in all languages.

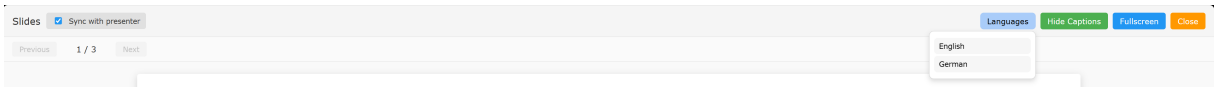


Figure 6: Slide viewer interface with multilingual navigation options. Users can switch between languages, browse slides independently of the presenter through an out-of-sync mode, and subsequently use the sync toggle to realign with the live presentation.

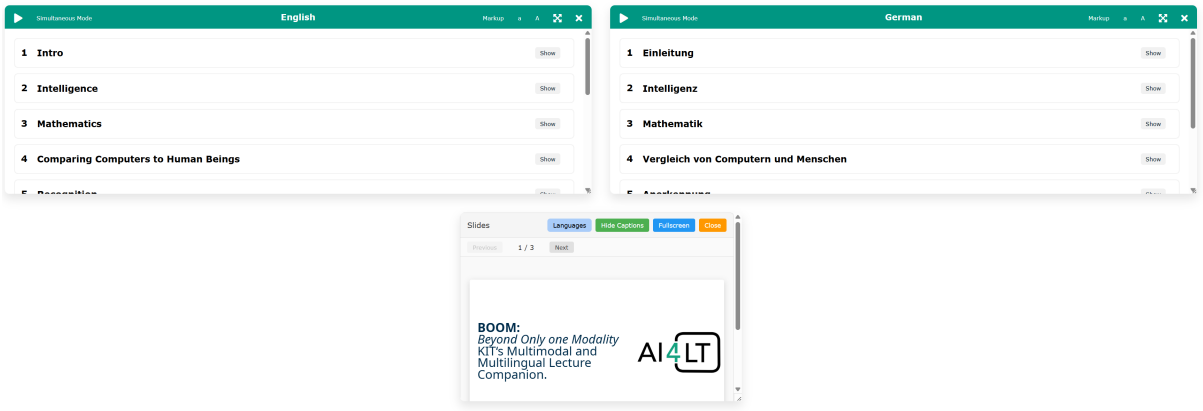


Figure 7: The interface also allows to see the translations in multiple languages along with the current slide.

Figure 8: Participant full screen view of the slide interface showing slides with caption overlay in German.