Re:Member: Emotional Question Generation from Personal Memories

Zackary Rackauckas[1], Nobuaki Minematsu[2] Julia Hirschberg[1],
[1]Columbia University, [2]The University of Tokyo,
zcr2105@columbia.edu, mine@gavo.t.u-tokyo.ac.jp, julia@cs.columbia.edu

Abstract

We present Re:Member, a system that explores how emotionally expressive, memory-grounded interaction can support more engaging second language (L2) learning. By drawing on users' personal videos and generating stylized spoken questions in the target language, Re:Member is designed to encourage affective recall and conversational engagement. The system aligns emotional tone with visual context, using expressive speech styles such as whispers or late-night tones to evoke specific moods. It combines WhisperX-based transcript alignment, 3-frame visual sampling, and Style-BERT-VITS2 for emotional synthesis within a modular generation pipeline. Designed as a stylized interaction probe, Re:Member highlights the role of affect and personal media in learner-centered educational technologies.

1 Introduction

As language learning technologies evolve, there is growing interest in systems that go beyond rote vocabulary drills or disembodied text. Research in Human–Computer Interaction (HCI) and Natural Language Processing (NLP) has shown that social presence, emotional involvement, and personal relevance significantly improve learning outcomes, especially for the acquisition of second languages (L2). However, most existing tools are based on generic or de-contextualized content, limiting their potential to tap into the emotional and mnemonic power of a learner's lived experiences.

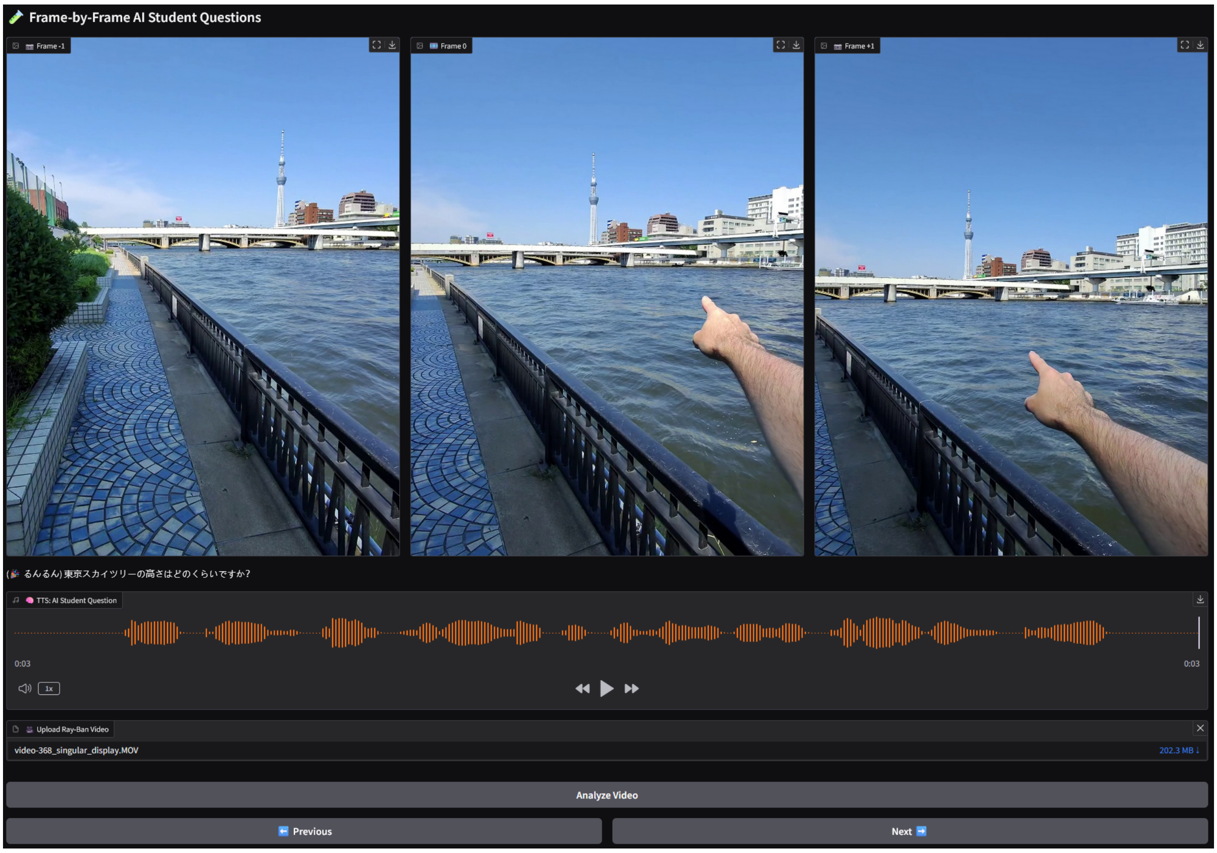
A growing body of HCI research has explored how large language models (LLMs) can support learners and designers through agent-assisted creativity. Systems such as IdeationWeb (Shen et al., 2025) and Promptify (Brade et al., 2023) scaffold user interaction with generative models, enabling iterative refinement and analogical exploration. In language learning, voiced chatbot interfaces, such as conversational characters and stress-free conversational partners (Rackauckas and Hirschberg, 2025b; Aiba et al., 2024), have shown how conversational systems can support learners by tailoring responses to their needs. Related work in agent-assisted creativity and co-design highlights the importance of aligning model outputs with user intent and emotional framework (Shaer et al., 2024; Sun et al., 2025).

From an NLP perspective, recent work on question generation has moved toward more context-sensitive and user-aligned output. Newer methods leverage LLMs for conversational foresight (Guo et al., 2024) and empathetic dialogue (Siyan et al., 2024) where the user's inferred state shapes responses for empathy and engagement (Rashkin et al., 2019). Our work contributes to this work by combining environmental-aware inference from sequential visual frames with LLM-based question generation in a real-time learner interface. The system supports reflective learning by surfacing system-generated, context-sensitive questions that adapt to the learner's evolving affective and attentional state, a goal aligned with broader calls for emotionally intelligent educational technologies (Darling-Hammond et al., 2017). This bridges recent work in HCI and NLP on responsive, learner-aware systems for mixed-initiative interaction.

Our system builds on previous work by grounding LLM-generated questions in video-based emotion cues, enabling emotionally responsive interactions that match the learner's current context. Specifically, we present Re:Member[1], an open-source system that turns videos of personal memories, also known as episodes, such as casual recordings of travel, family, or everyday life, into emotionally voiced, interactive prompts for language learning. By combining recent advances in large language models (LLMs), expressive speech synthesis,

[1] https://github.com/zackrack/Re-Member

Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP), pages 163–168
November 9, 2025 ©2025 Association for Computational Linguistics



(🎉 るんるん) 東京スカイツリーの高さはどのくらいですか?

video-368_singular_display.MOV

Figure 1: Example interface frame from video (1), showing (from top to bottom) three frames of sequential visual context, the generated emotion, the generated system question, a playable text-to-speech box, the name of the video file, the "Analyze Video" button, and "Previous" and "Next" buttons to navigate between sequential moments.

and vision-language processing, Re:Member analyzes short user-uploaded videos, extracts scene-relevant transcripts and images, and generates stylized spoken questions in the learner's target language. These questions are voiced in emotional speaking styles (e.g., playful, whispered, drowsy), selected to match the tone and atmosphere of the scene.

The core idea behind Re:Member is that emotionally salient, personally meaningful content may create deeper engagement for language learners, especially when paired with stylized voice output that mimics familiar social dynamics (e.g., a whisper from a friend or an excited exclamation). Rather than relying on synthetic neutrality, our system embraces affective richness as an instructional tool.

This paper introduces the design of Re:Member and demonstrates its capabilities as an emotional question-generation companion. We detail our architecture, design rationale, and sample outputs, and reflect on the broader implications for language education, affective computing, and memory-centered interaction.

2 System Overview

The goal of Re:Member is to generate emotionally expressive questions from personal memory videos for L2 (second language) conversational practice. These questions are designed to support language learning by connecting spoken language, visual content, and emotional speech.

2.1 Audio-Visual Segmentation

Given a video, we first extract its audio and apply voice activity detection (VAD) using the Silero VAD model (Team, 2024). This produces a list of speech segments, which we merge if the intervening silence is shorter than a 0.7 second threshold. Each segment is then transcribed using WhisperX (Bain et al., 2023), which produces high-quality transcripts along with accurate word-level timing alignment. This allows us to preserve the temporal correspondence between the transcript and the visual context.

2.2 Frame Sampling and Visual Context

To provide visual grounding for each spoken segment, we extract a 3-frame window per segment: one frame before, during, and after the midpoint of the segment. This is done using OpenCV (OpenCV contributors, 2025), and frames are resized and saved in a consistent format. The use of three temporally adjacent frames provides richer context than a single image and allows the language model to infer scene dynamics (e.g., motion, transitions, or emotional shifts).

2.3 Multimodal Question Generation

For each segment, we generate a natural Japanese-language question using GPT-4o (OpenAI et al., 2024), conditioned on both the transcript and the associated video frames. Frames are provided with the transcript segment. We instruct the language model to simulate the behavior of a friendly, curious learner asking questions to the person who filmed the video (see Appendix A). This encourages open-ended questions that are personally meaningful and draw emotional context primarily from the user's environment and accompanying speech content.

2.4 Emotion Style Selection

To enhance engagement and match the emotional tone of each moment, we generate a corresponding speaking style label from a fixed set of Japanese emotional styles:

1. るんるん (cheerful),
2. ささやきA（無声）(silent whisper),
3. ささやきB（有声）(voiced whisper),
4. ノーマル (neutral),
5. よふかし (late-night relaxed).

We choose these styles because they align with the expressive capabilities of the pre-trained TTS model used in the next section. The language model is instructed to choose an option from this list that matches the mood and context of the visual scene (see Appendix A). To encourage variation, we set the temperature to 1 and maintain a short history of recent emotion labels. If the generated style matches any of the last two used, the model is re-queried up to five times. This re-query mechanism helps prevent repetition and promotes emotional diversity across segments. The selected emotion is then passed to the speech synthesis stage.

2.5 Expressive Speech Synthesis

The generated question and selected style label are sent to a local Style-BERT-VITS2 model (litagin02, 2024) for emotionally expressive Japanese text-to-speech synthesis. Specifically, we use a model trained from the Ami Koharune UTAU voicebank (Amitaro, 2025). This model supports fine-grained style control via natural language emotion labels and produces speech that reflects not only the content of the question, but also its mood and delivery (Rackauckas and Hirschberg, 2025a). The result is an audio clip paired with the original frames and transcript, allowing for emotionally aligned language learning experiences.

2.6 Interactive User Interface

Users can upload videos and browse the resulting questions in a Gradio (Abid et al., 2019) interface with synchronized:

1. Three representative frames per segment,
2. The generated Japanese question and emotion text,
3. Emotionally styled speech playback.

This interface enables learners to engage with their own personal content in an emotionally aware way, making the experience more memorable and contextually grounded.

3 Illustrative Outputs

We demonstrate the system with two sample videos: (1) A video of a walk along Tokyo's Sumida River with the commentary playing the role of a language teacher, and (2) a video of the user boarding a train in Japan with spoken instructions for boarding the train. Both videos were recorded with Meta Ray-Ban Glasses, and (1) is 1 minute and 31 seconds in length while (2) is 31 seconds in length. For video (1), the system segmented and analyzed 13 moments, generating an emotion, a student question, and text-to-speech for each moment.

For each of the 13 segments in video (1), the system generated a natural Japanese-language question grounded in both the visual scene and the transcript. These questions reflect a consistent student-like curiosity, such as asking what kinds of boats travel through the river or how tall the Tokyo Skytree is. The selected emotion styles were well-matched to the riverfront setting, with a majority in the gentle voiced whisper style, interspersed with more upbeat cheerful and late-night relaxed tones. All five available emotion styles appeared at least once, showing that the variation mechanism functioned appropriately given the consistent environment. The visual frames used as context were sampled from before, during, and after each utterance, helping the language model infer motion and visual focus —- such as when the user points to a boat or approaches a bridge. Each segment resulted in synchronized audio narration with emotional speech, allowing for immersive and pedagogically meaningful playback. A select moment from video (1), as seen in Figure 1 shows the user pointing their finger to Tokyo Sky Tree, a tall tower on the other side of the river. For this moment, the system generated the question

> 東京スカイツリーの高さはどのくらいですか？
> **Translation:** About how tall is Tokyo Sky Tree?

with the cheerful emotion (るんるん).

For video (2), the system identified and processed three distinct segments, each aligned with the user's spoken instructions for boarding a train in Japan. The generated questions reflect a student-like curiosity about practical aspects of the scene, such as the convenience of using trains near event venues or the layout of the train interior. Emotion styles were chosen to match the focused, informational tone of the video: a balance of voiced whisper, silent whisper, and neutral speech was used across the three questions. Though the short duration of the video limited the range of styles, the variation mechanism successfully avoided repetition and produced a tone consistent with the setting. Visual context was drawn from three-frame windows centered on each utterance, allowing the language model to reference specific spatial cues – such as when the user physically steps onto the train. As with video (1), the result is synchronized emotional narration paired with visually grounded, pedagogically meaningful questions. The generated questions and associated emotion styles are shown below:

> **Question:** (Silent whisper) 試合が行われている場所での電車の利用はどのように便利ですか？
> **Translation:** How is using the train convenient near where the event is being held?
> **Question:** (Neutral) この電車の車内はどのように見えますか？
> **Translation:** What does the inside of this train look like?
> **Question:** (Voiced whisper)この電車の車両には特別な座席やスペースがありますか？
> **Translation:** Does this train car have special seats or areas?

4 Discussion and Future Work

By using a learner's own memory videos as input, Re:Member creates interactions in which the learner appears as the main character rather than a passive observer. Unlike textbook stories, these moments are drawn from the learner's real experiences, ensuring strong personal relevance and evoking the raw, multimodal sensations originally felt — the sights, sounds, and emotions of the scene. Such vivid, embodied memories form a powerful substrate for retaining new linguistic forms, especially when voiced through Re:Member's expressive speech synthesis that mirrors the affect of the original experience. While the current implementation targets Japanese, the pipeline generalizes to other language settings where learner identity and emotional relevance shape engagement. Future work may explore adaptive selection of emotion styles, more nuanced alignment between visual and emotional cues, and interactive control over style and content. Longitudinal deployments could evaluate how learners interact with memory-grounded prompts over time and whether affectively voiced questioning measurably enhances learning, including validation of emotion–scene alignment.

Limitations

Re:Member assumes clean, monolingual speech from a primary speaker, and performance may degrade in the presence of overlapping dialogue, background noise, or multilingual utterances. Emotion style selection is based on LLM prompting rather than perceptual modeling and may at times produce mismatched or overly expressive styles. The system has not yet been evaluated with users; it is presented as a design and technical demonstration. Finally, as it operates on personal memory videos, future iterations must consider consent, emotional safety, and data privacy, for example, by supporting local-only processing and explicit opt-in use of autobiographical media.

References

Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Mayuko Aiba, Daisuke Saito, and Nobuaki Minematsu. 2024. A chatgpt-based oral q&a practice system for first-time student participants in international conferences. In *Interspeech 2024*, pages 5202–5203.

Amitaro. 2025. Ami koharune utau voicebanks. https://amitaro.net/utau/en_ongen-list.html. Version info available for each bank; accessed August 2025.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-image generation through interactive prompt exploration with large language models. *Preprint*, arXiv:2304.09337.

Linda Darling-Hammond, Maria E. Hyler, and Madelyn Gardner. 2017. Effective teacher professional development. Technical report, Learning Policy Institute, Palo Alto, CA.

Shasha Guo, Lizi Liao, Jing Zhang, Cuiping Li, and Hong Chen. 2024. PCQPR: Proactive conversational question planning with reflection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11266–11278, Miami, Florida, USA. Association for Computational Linguistics.

litagin02. 2024. Style-bert-vits2. https://github.com/litagin02/Style-Bert-VITS2. Accessed: 2025-01-22.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenCV contributors. 2025. OpenCV: Open Source Computer Vision Library. https://github.com/opencv/opencv. Version 4.12.0, released July 2, 2025.

Zackary Rackauckas and Julia Hirschberg. 2025a. Benchmarking expressive japanese character text-to-speech with vits and style-bert-vits2. *Preprint*, arXiv:2505.17320.

Zackary Rackauckas and Julia Hirschberg. 2025b. Learning japanese with jouzu: Interaction outcomes with stylized dialogue fictional agents. *Preprint*, arXiv:2507.06483.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Orit Shaer, Angelora Cooper, Andrew L. Kun, and Osnat Mokryn. 2024. Toward enhancing ideation through collaborative group-ai brainwriting. In *Joint Proceedings of the ACM IUI Workshops 2024*, Greenville, South Carolina, USA. CEUR-WS.org. March 18–21, 2024.

Hanshu Shen, Lyukesheng Shen, Wenqi Wu, and Kejun Zhang. 2025. Ideationweb: Tracking the evolution of design ideas in human-ai co-creation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Li Siyan, Teresa Shao, Zhou Yu, and Julia Hirschberg. 2024. EDEN: Empathetic dialogues for English learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3492–3511, Miami, Florida, USA. Association for Computational Linguistics.

Fuze Sun, Lingyu Li, Shixiangyue Meng, Xiaoming Teng, Terry Payne, and Paul Craig. 2025. Integrating emotional intelligence, memory architecture, and gestures to achieve empathetic humanoid robot interaction in an educational setting. *Preprint*, arXiv:2505.19803.

Silero Team. 2024. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier.

A Appendix A

A.1 Question Generation Prompt

For question generation, we give the LLM the following prompt. English translations are included for clarity only and are not shown to the model.

あなたは英語を学んでいる好奇心旺盛でフレンドリーな学生です。先生が作ったビデオを見て学んでいます。

```
You are a curious and friendly student
learning English. You are watching a
video made by your teacher.
```

映像のシーンと先生が話している内容の両方を考慮してください。

```
Take both the visual scene and what the
teacher is saying into account.
```

学習を深めるために、一つ短く関連性の高い質問をしてください。

```
Ask one short, highly relevant question
to deepen your learning.
```

画像に写っている人を特定したり、身元を推測したり、名前に言及したりしないでください。
Do not identify or guess the identity of anyone in the image, and do not refer to names.
年齢、性別、身元、名前についての推測を避けてください。
Avoid guessing age, gender, identity, or names.
質問のみを返し、それ以外は返さないでください。
Only return the question and nothing else.
必ず日本語で質問をしてください。
Make sure to ask the question in Japanese.

A.2 Emotion Selection Prompt

For selecting emotions in the context of the scene, we give the LLM the following prompt. English translations are included for clarity only and are not shown to the model.

あなたは感情ラベル分類機です。以下の5つのラベルの中から **1つだけ** を選んで日本語で出力してください：
You are an emotion label classifier. Select and output **only one** label in Japanese from the five options below:
1. るんるん
1. Runrun (cheerful or bubbly tone)
2. ささやきA（無声）
2. Whisper A (voiceless whisper)
3. ささやきB（有声）
3. Whisper B (voiced whisper)
4. ノーマル
4. Normal
5. よふかし
5. Late-night (sleepy or relaxed nighttime tone)

出力は **上記の5つのラベルのいずれか1つだけ** にしてください。
Your output must be exactly one of the five labels listed above.
絶対に説明・理由・挨拶・謝罪などを含めてはいけません。
Do not include any explanation, reasoning, greetings, or apologies under any circumstances.
他のテキストを含んだら重大なフォーマットエラーです。
Including any other text is a serious formatting error.
視覚的な背景（画像）とセリフの両方を考慮して、最も表現豊かで印象に残るスタイルを優先してください。
Prioritize the most expressive and memorable style by considering both the visual background (image) and the spoken dialogue.
同じスタイルばかり繰り返すことを避けてください。
Avoid repeatedly selecting the same style.
「ノーマル」は控えめにし、場面に応じて他のスタイルを積極的に使ってください。
Use “Normal” sparingly, and actively choose other styles based on the scene.