Archival Paper

CVPR JUNE 3-7, 2026 DENVER COLORADO

MES-Bench: A Benchmark for Multimodal Elaborative Simplification and Comprehensibility Evaluation in Language Learning

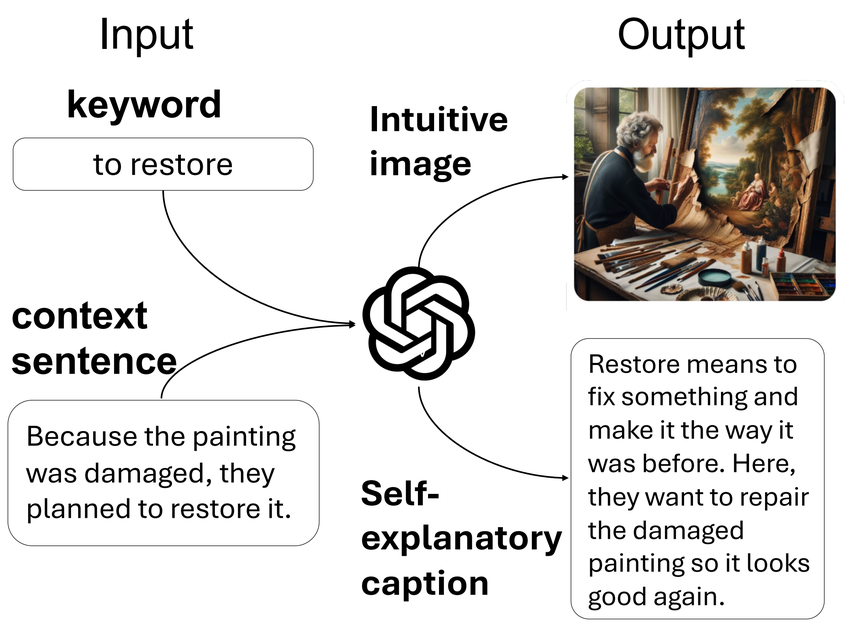
Martyna Gruszka[1], Risa Shinoda[2], Taiki Miyanishi[2], Takumi Hirose[1], Nakamasa Inoue[1]
Institute of Science Tokyo[1], The University of Tokyo[2]

Motivation

- Language learning benefits from **comprehensible input** (understanding through context)
- Elaborative simplification adds explanations, but is mostly **text-only**
- Multimodal models can generate images + text, but:

→ **What makes them truly understandable is unclear**
→ **Can models evaluate this reliably?**

Multimodal Elaborative Simplification



Generate image + caption to explain a target keyword in context

Benchmark: MES-Bench

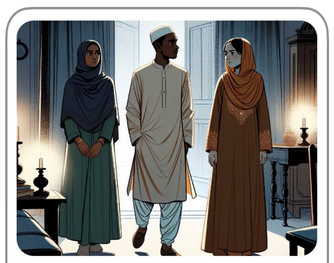
Evaluation of multimodal explanations
Pairwise comparison Plain vs Rich
Ratings + Preference + Rationales

Do multimodal explanations actually help learners understand?

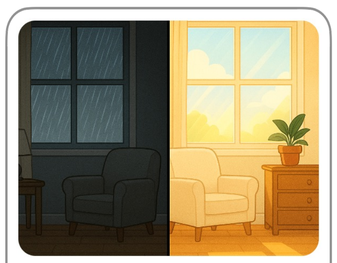
Restore means to fix something and make it the way it was before. Here, they want to repair the damaged painting so it looks good again.

Restore means to repair something and bring it back to its original condition. In the image, the old painting on the left has been restored to the fresh version shown on the right.

to restore (B2)



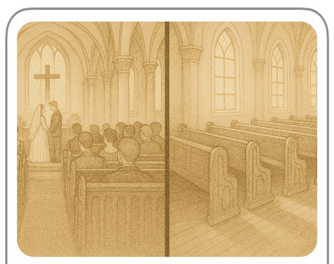
Sombre means dark, gloomy, or serious. In the context of the room, it suggests the room felt serious or had a dark, heavy atmosphere.



Sombre means dark and serious in mood – not bright or cheerful. For example, the room on the left feels sombre, dark, and a bit sad, while the one on the right looks bright and cheerful instead.

sombre (C1)

A pew is a long bench with a back, found in churches, whee people sit during a service or ceremony.



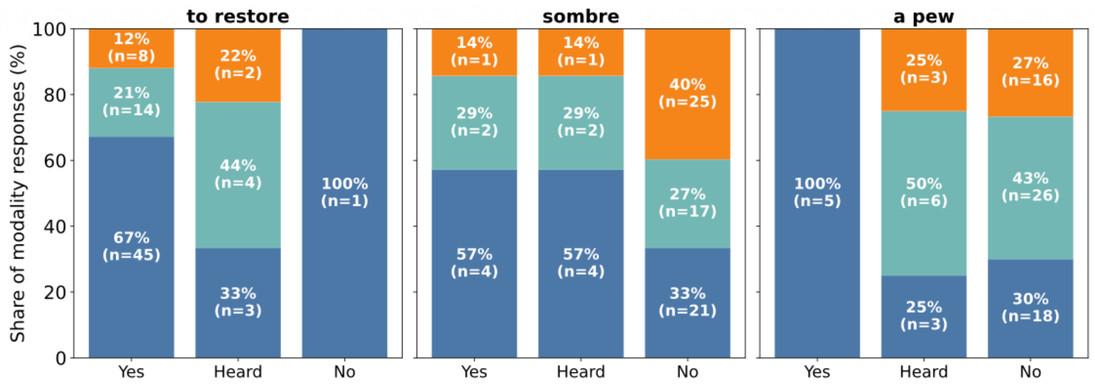
The people in the image are seated in pews. **A pew** is a long bench in a church where people sit during a service or ceremony – like the ones shown on the right.

a pew (C1)

Results

Keyword	Flashcard	Pref. H	Pref. M	Mean H	Mean M	Median H	Median M	Std. H	Std. M
to restore	Plain	27	0	3.36	3.69	3.0	4.0	1.15	0.48
	Rich	47	16	3.99	5.00	4.0	5.0	1.03	0.00
sombre	Plain	35	0	3.61	3.50	4.0	3.5	1.13	0.52
	Rich	36	16	3.51	4.94	4.0	5.0	1.20	0.25
a pew	Plain	34	5	3.99	4.31	4.0	4.0	1.21	0.48
	Rich	21	11	3.69	4.56	4.0	5.0	1.17	0.73

Modality preference by familiarity across keywords



Key insights

Humans	MLLMs
• Prefer conciseness • Sensitive to ambiguity • Focus on realism	• Prefer descriptiveness • Sensitive to contrast • Focus on layout

Learner preferences are not uniform

- **Beginners**: people in the image + contrastive layouts
- **Advanced**: simple + object-focused images

Key factors influencing flashcard comprehensibility
(Human-centered evaluation criteria)

- Explanation Brevity Conciseness
- Image-text Alignment Referential clarity
- Presence of concrete examples
- Placement of main explanation
- Paraphrasing Lexical support
- Realism Image quality
- Lighting Visibility
- Level of detaiil
- Focus Perspective
- Presence of visual cues
- Layout structure

Main contributions

- First benchmark for MES
- Humans ≠ MLLMs in evaluating comprehension
- Proposed human-centered evaluation rubric