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

Anonymous CVPR submission

Paper ID 12

Abstract

Recently, alternatives to traditional translation-based language learning methods have gained increasing attention, particularly the Natural Approach, which emphasizes content that learners can understand through context rather than translation. This perspective motivates methods that make meaning inferable through contextual support, including elaborative text simplification, where clarifying information is added to improve comprehension. However, existing work has remained largely text-only, and while current Multimodal Large Language Models (MLLMs) can generate high-quality images, making multimodal image-caption elaborations increasingly feasible, such images are often overly generic and fail to provide the specific contextual cues needed to support the intended cognitive associations. Moreover, as such materials become easier to generate at scale, it remains unclear whether MLLMs can reliably evaluate their pedagogical comprehensibility. To study this problem, we introduce ***MES-Bench****, a benchmark for evaluating the Multimodal Elaborative Simplification (MES) task, that enables direct comparison between human and MLLM judgments through pairwise evaluation with both quantitative and qualitative feedback. Using this framework, we perform a human-MLLM study in which both groups evaluate the same flashcard-style elaboration pairs. The results show that their comprehensibility judgments are not interchangeable: models tend to prefer contrast-heavy designs, whereas humans are more sensitive to visual focus and contextual grounding. Motivated by these gaps, we derive a human-centered rubric for evaluating multimodal elaborations in language learning.*

1. Introduction

Language learning research has long argued that effective instruction need not rely on translation or explicit grammar explanation, but can instead be based on *comprehensible input*, that is, language learners can understand through context [22–24, 34]. This view is central to the *Natural Approach*, which proposes that acquisition develops through understanding messages rather than through translation-based study [25]. Reading is a practical way to provide comprehensible input, giving learners sustained exposure to understandable language at their own pace. Prior work shows that meaningful, extensive, and self-selected reading can support second-language development and practical proficiency [33, 35, 36], especially when texts are interesting and matched to learner needs [14]. This perspective motivates methods that make meaning inferable through context rather than translation. In Natural Language Processing (NLP), a closely related idea appears in elaborative text simplification, where the goal is not only to reduce difficulty but also to improve comprehension by adding clarifying information [46, 48]. So far, however, this line of work has focused primarily on text alone. More broadly, recent studies have also explored text-based educational applications such as personalized multiple-choice question generation [20], student learning behavior simulation [50], and pedagogy-oriented tutoring [13]. At

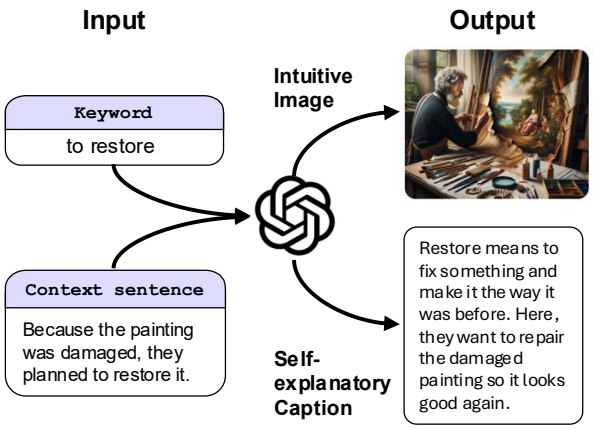

Figure 1. **Overview of Multimodal Elaborative Simplification (MES).** Given a target keyword and a context sentence, the MES task generates an image-caption pair consisting of an intuitive image and a self-explanatory caption, so that the learner can infer the meaning of the target keyword from multimodal context without translation.

the same time, emerging work has trained MLLMs to perform reasoning in multimodal contexts [31, 32], opening the way for contextual support beyond text alone. Recent applications include multimodal feedback generation [28], evaluation of hand-drawn mathematical understanding [5], and personalized learning support [26]. In language learning more specifically, MLLMs are beginning to be used for adaptive multimodal vocabulary flashcard generation [19]. Flashcards are a natural testbed for this setting because they remain a core tool for vocabulary acquisition, while digital flashcards have been shown to improve retention [49] and motivation [39].

As such materials become easier to generate at scale, the key question is not only whether they are fluent or visually appealing, but whether they are genuinely comprehensible and pedagogically effective. This is especially important because MLLM-generated materials are increasingly evaluated by models themselves, even though prior work has found fairness-related distortions [47], judgment bias in human-vs-LLM settings [9], self-preference effects [29], and positional or stylistic bias in pairwise evaluation [51]. In multimodal settings, models may also over-reward surface image-text alignment [8], despite the need for more fine-grained evaluation [27], and may attend to visual details differently from humans [12]. These concerns are particularly serious in education, where ethical and pedagogical risks of LLM use have already been noted [21, 37].

To support language learning and test whether MLLMs can reliably evaluate pedagogical comprehensibility, we introduce **MES-Bench**, a benchmark for *Multimodal Elaborative Simplification* (MES), where a target phrase in context is explained through a flashcard-style image-caption elaboration supporting translation-free meaning inference inspired by the Natural Approach. MES-Bench enables direct comparison between human and MLLM judgments through pairwise evaluation with quantitative and qualitative feedback. Using this framework, 77 participants and 16 MLLMs evaluate the same flashcard pairs, showing that their comprehensibility judgments are not interchangeable: models tend to prefer contrast-heavy designs, whereas humans are more sensitive to visual focus and contextual grounding. Based on these gaps, we derive a human-centered rubric for evaluating multimodal elaborations.

2. Related work

A closely related line of research is Text Simplification (TS), which aims to make complex content more accessible by reducing linguistic complexity or paraphrasing across difficulty levels [6, 10], and has been shown to improve comprehension and reading ease [15, 18]. MES is particularly related to elaborative simplification, where comprehension may require adding useful information rather than merely compressing text [46, 48]. Our task extends this idea beyond text by asking how such contextual support can be distributed across modalities. Recent document-level work has extended simplification toward multimodal outputs. Liu et al. [30] combine document simplification with automatic illustration generation, while Souayed et al. [45] study structured prompting for cognitively accessible images from simplified texts. Our work builds on this direction, but shifts the focus from generation to evaluation: rather than asking only how to produce multimodal explanations, we ask how to assess whether they are actually comprehensible to learners.

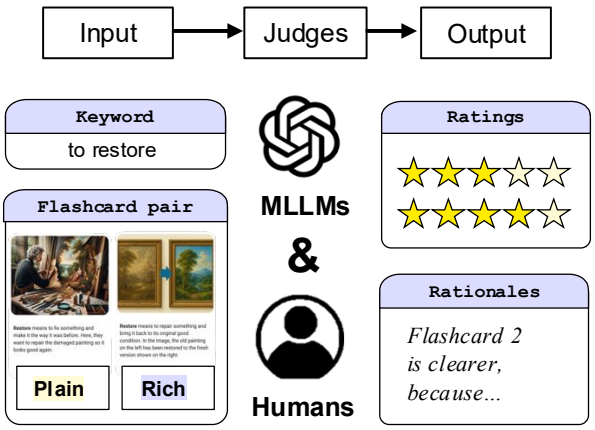

Figure 2. **Overview of MES-Bench.** MES-Bench is a benchmark for evaluating comprehensibility in the MES task for language learning.

3. Multimodal Elaborative Simplification

We define *Multimodal Elaborative Simplification* (MES) as a multimodal generation task for translation-free vocabulary explanation. The goal is to support intuitive meaning inference by combining textual and visual context. A MES instance provides (i) a **target keyword** w (a word to be learned) and (ii) a **context sentence** s in which w appears (to disambiguate its sense). As illustrated in Figure 1, these two inputs are used to generate an **image-caption pair** $f = (I, c)$, where I is intended to be an **intuitive image** and c a **self-explanatory caption**. The image should visually ground the meaning of the target keyword, while the caption should explain it in simple contextualized language. Together, (s, I, c) should allow a learner to infer the meaning of w without relying on external knowledge or translation. In our formulation, the target keyword w is explicitly provided to make the intended learning objective clear for both generators and judges.

3.1. MES-Bench

MES-Bench is a benchmark for evaluating the comprehensibility of the MES task in the language learning domain. As shown in Figure 2, the input consists of a target keyword

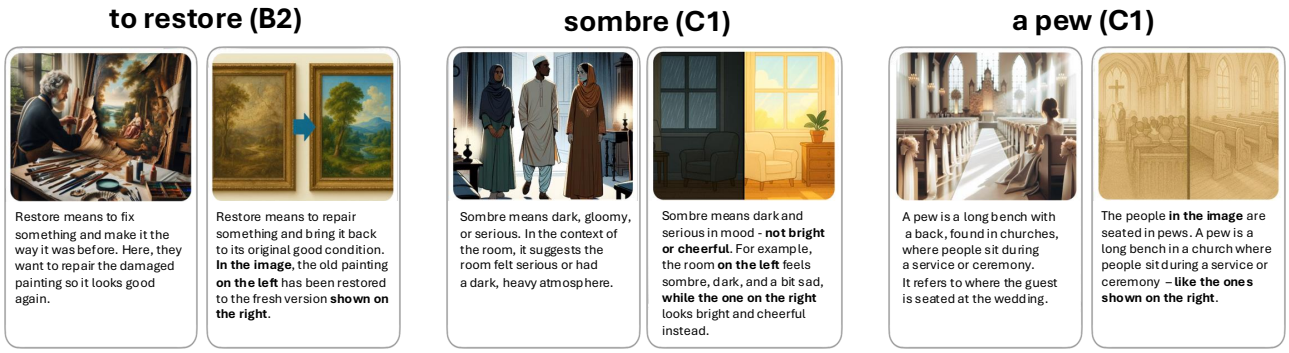

Figure 3. **Flashcard pairs used in MES-Bench across three target keywords: the verb *to restore*, the adjective *sombre* and the noun *a pew*.** For each keyword, the left flashcard shows the *Plain* version, while the right flashcard represents the *Rich* version.

Basic Instructions for Caption Generation

```
Write a short explanation for the keyword "{}" based
on the following text: "{}". The explanation should
be simple, clear, and self-contained — it should
make sense even without translation or external
knowledge.
```

Basic Instructions for Image Generation

```
Generate a realistic image that clearly and
naturally represents the meaning of the keyword "{}"
in the context of the following text: "{}".Visually
highlight the meaning of the keyword by depicting
its most common function, appearance, or interaction
as it would occur in real life within the given
context. The image should be clear, realistic,
and easy to understand without any text labels.
```

Figure 4. Basic prompting instructions for caption and image generation in MES-Bench.

and a **flashcard pair** containing two candidate explanations, *Plain* and *Rich*, for the same item. These candidates are then assessed by **both MLLMs and human judges**. The output comprises two forms of judgment: **ratings**, which quantify comprehensibility, and **rationales**, which explain the preference in natural language (e.g., why one flashcard is clearer than the other). In this way, each benchmark item stores (w, s) and optional metadata (e.g., CEFR level, part of speech), together with side-by-side candidate flashcards that support pairwise comparative evaluation with both quantitative and qualitative feedback.

3.2. Flashcard candidates

Figure 3 presents the flashcard pairs constructed for this study, covering three target items across two CEFR proficiency levels.[1] The figure includes one verb (*to restore*, B2), one adjective (*sombre*, C1), and one noun (*a pew*, C1), allowing comparison across different parts of speech as well as across items that may vary in expected learner familiarity, with the B2 item assumed to be relatively more familiar than the C1 items. For each target item, we created two candidate flashcards: a *Plain* version on the left and a *Rich* version on the right. Both flashcards combine an image with a short explanatory caption, but they use different ways to help learners infer the meaning. Given a target keyword and a short context sentence (for sense disambiguation), we prompted GPT-4o [44] to generate the elaborative caption and DALL·E 3 [41] to generate the corresponding image.

Plain Flashcard. The *Plain* flashcard is designed to be direct and minimal. Figure 4 presents the basic prompt instructions for caption and image generation, designed to produce a short, clear, self-contained explanation understandable without translation or external knowledge, together with a clear and realistic image highlighting the keyword's most common function, appearance, or real-world use.

Rich Flashcard. In contrast, the Rich flashcard is designed to provide stronger contextual elaboration and more explicit guidance for connecting the visual and textual information. Its caption prompt encouraged explanatory strategies, such as the use of examples and antonyms. Its image prompt also varied the visual strategy and style, for example through process-versus-outcome depictions, contrastive cues, or multi-view scenes. These variations were introduced to test whether such design differences affect judge preferences. As shown in Figure 3, this added support takes different forms depending on the target item, such as before-after comparison for *to restore*, contrastive scenes for *sombre*, or explicit reference to the relevant object or region in the image for *a pew*.

4. Experiments

4.1. Human participants

We collected responses from 77 human participants, whose age distribution was heavily skewed toward younger individuals: 90.9% were aged 18-24. A small proportion were

[1] CEFR (Common European Framework of Reference for Languages) defines proficiency bands: A1-A2 (beginner), B1-B2 (intermediate), and C1-C2 (advanced).

Keyword	Flashcard	Pref.		Mean		Median		Std.	
		H	M	H	M	H	M	H	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

Table 1. **Human (H) and MLLM (M) ratings for each flashcard pair (Plain vs. Rich)**. 'Pref.' indicates the number of evaluators favoring each flashcard; remaining human responses, are ties.

MLLM	to restore		sombre		a pew	
	Plain	Rich	Plain	Rich	Plain	Rich
Claude 4.5 Haiku [1]	4.0	5.0	3.0	4.0	4.0	5.0
Claude Sonnet 4.5 [2]	3.0	5.0	3.0	5.0	4.0	5.0
GPT-4o [44]	4.0	5.0	4.0	5.0	5.0	4.0
GPT-5 [43]	4.0	5.0	4.0	5.0	4.0	5.0
GPT-5.1 [42]	4.0	5.0	4.0	5.0	5.0	4.0
Gemini 2.5 Flash [11]	4.0	5.0	3.0	5.0	4.0	5.0
Gemini 2.5 Pro [11]	4.0	5.0	4.0	5.0	5.0	5.0
Gemini 3 Pro [17]	3.0	5.0	3.0	5.0	4.0	5.0
Gemma 3 (27B) [16]	4.0	5.0	3.0	5.0	4.0	5.0
LLama 4 Scout [7]	4.0	5.0	4.0	5.0	4.0	5.0
LLama 4 Maverick [7]	4.0	5.0	3.0	5.0	5.0	4.0
Mistral Small 3.1 (24B) [38]	4.0	5.0	4.0	5.0	4.0	5.0
Nemotron Nano V2 VL (12B) [40]	3.0	5.0	3.0	5.0	4.0	5.0
Qwen2.5VL (32B) [3]	4.0	5.0	4.0	5.0	4.0	3.0
Qwen3 VL Thinking (8B) [4]	3.0	5.0	3.0	5.0	5.0	3.0
Qwen3VL (8B) [4]	3.0	5.0	4.0	5.0	4.0	5.0

Table 2. **Per-MLLM ratings** for *Plain* flashcard and *Rich* flashcard variants across keywords.

aged 45 or older (5.2%), and only 3.9% fell into the 25-34 age range. The sample was also predominantly composed of lower- to mid-intermediate learners, with relatively few participants at advanced proficiency levels. The largest group consisted of B1 learners, representing 40.3% of the sample. This was followed by B2 learners at 20.8%. A2 participants accounted for 16.9%, while 14.3% reported A1 proficiency. The smallest group comprised C1 learners, making up 7.8% of respondents.

4.2. Human and MLLM judging protocol

To compare human and MLLM perceptions of flashcard comprehensibility, we evaluated the same Plain-Rich flashcard pairs for three target keywords under a matched setup. For each keyword, evaluators first saw the target word in a sample sentence and then assessed the pair. Human participants reported keyword familiarity, rated each flashcard on a 1-5 scale, selected the clearer flashcard, provided a brief justification, and indicated whether understanding came mainly from the image, the caption, both equally, or neither. In parallel, 16 MLLMs evaluated the same pairs under a structured prompt, producing two 1-5 scores, a preferred flashcard, and a short justification.

Group	n	to restore		sombre		a pew	
		Δ	p-value	Δ	p-value	Δ	p-value
Human (All)	77	0.623	**0.006**	-0.104	0.590	-0.299	0.115
Human (A1)	11	0.364	0.533	1.091	0.100	-0.636	0.363
Human (A2)	13	0.308	0.631	-0.923	0.142	-0.077	0.945
Human (B1)	31	0.710	0.071	0.000	0.991	-0.355	0.170
Human (B2)	16	0.500	0.281	-0.438	0.243	-0.313	0.523
Human (C1)	6	1.667	**0.031**	-0.167	1.000	0.167	1.000
MLLM (All)	16	1.313	**0.000**	1.438	**0.000**	0.250	0.346

Table 3. **Paired comparison results by group.** Δ denotes the mean score difference between the Rich and Plain flashcards. Bold indicates statistically significant p-values ($p < 0.05$). Delta cells are color-coded: purple for $\Delta > 0$ (Rich Flashcard favored) and yellow for $\Delta < 0$ (Plain Flashcard favored).

4.3. Results

We counted per-keyword preferences for the Plain and Rich flashcards, and computed descriptive statistics (mean, median, spread) for both human and MLLM ratings. We also analyzed preference proportions and tested paired differences using Wilcoxon signed-rank tests across cohorts. Although humans and MLLMs partially agree overall, they diverge on specific keywords. Table 1 reports overall human and MLLM rating statistics, and Table 2 details per-model scores for each flashcard pair.

Flashcard Preference Flashcard preferences show partial agreement between humans and MLLMs, but also clear keyword-specific divergences. For *to restore*, humans more often preferred the Rich flashcard (61.0%) than the Plain one (35.1%), with 3.9% ties, while all MLLMs preferred Rich (100%). For *sombre*, human preferences were nearly balanced (Rich: 46.8%, Plain: 45.5%, ties: 7.8%), whereas all MLLMs again preferred Rich (100%). For *a pew*, humans more often preferred Plain (44.2%) than Rich (27.3%), with a relatively large proportion of ties (28.6%), while MLLMs more often preferred Rich version (68.8%) than the Plain one (31.3%).

Ratings and Statistical Comparison MLLM standard deviations are consistently low, indicating tight inter-model agreement, whereas human standard deviations are larger across conditions, especially for the Plain flashcard of *to restore* and the Rich flashcard of *a pew*, reflecting greater variability in learner judgments. To quantify these differences, we tested preference for the Rich over the Plain flashcard using Wilcoxon signed-rank tests, as shown in Table 3. In the human data, only *to restore* is significant overall ($p = 0.006$), and by proficiency only the C1 group shows a significant effect for *to restore* ($p = 0.031$), indicating generally variable human preferences. In contrast, MLLMs show strong, consistent preferences for Rich flashcard on both *to restore* ($p < 0.001$) and *sombre* ($p < 0.001$).

Keyword	Human comments	MLLM rationales
to restore (Human: Rich ; MLLM: Rich)	"**The comparison before and after** restoration can make me better understand the meaning of this word."	"Flashcard 2 is more effective because **the "before and after" format** provides a very clear and direct visual comparison."
sombre (Human: Rich ; MLLM: Rich)	"We can recognize and imagine well because of the **existence of the opposite** image."	"Flashcard 2 is easier to understand because it provides **a clear visual contrast** (...)"
a pew (Human: Plain ; MLLM: Plain)	"Because attention is more easily drawn to the long bench **the woman in the image is sitting on.**"	"Flashcard 1 is more comprehensible because **it directly shows a person sitting on a pew**, making the object and its context immediately clear."
to restore (Human: Plain ; MLLM: Rich)	"The illustration is a little hard to see, but the English sentence is simple and **easy to understand**."	"...Flashcard 1's caption is **vague and lacks visual contrast**, making it harder to grasp the full meaning of "restore"
sombre (Human: Plain ; MLLM: Rich)	"**Including people with serious expressions** makes it easier to convey the nuance of a sombre environment..."	"...Flashcard 1 is accurate but more abstract and **distracted by people**, making the mood less immediately clear."
a pew (Human: Plain ; MLLM: Rich)	"...The other one does not seem to need a comparison-style layout; in fact, it feels like the **comparison gets in the way of interpretation**."	"Flashcard 2 is more effective because of its **split-panel design**, (...) showing its structure clearly without people or decorations obscuring it."

Table 4. **Representative cases of agreement and divergence between human participants and the MLLM in qualitative feedback.** Flashcard 1 corresponds to the "left" Plain card, and Flashcard 2 corresponds to the "right" Rich card. The labels in parentheses in each row indicate the card preferred by humans and by the MLLM.

5. Qualitative Analysis

To better understand the quantitative patterns and characterize the features humans and MLLMs appear to prioritize when judging comprehensibility, we analyzed participants' free-text explanations and compared them with MLLM rationales. For each keyword block, we first identified the flashcard favored in the ratings and then examined the accompanying comments.

Human-MLLM Agreement and Divergence Across keywords, several common themes emerged. Both humans and MLLMs often favored explanations that made the target meaning easier to infer by clearly linking the caption and the image. This pattern can be seen in the agreement cases shown in the top three rows of Table 4. For *to restore*, most human participants and all MLLMs agreed that before-and-after comparison made the meaning visually explicit. For *sombre*, human preferences were nearly balanced, but those who favored the Rich flashcard, along with all MLLMs, pointed to the usefulness of contrast between opposite moods. For *a pew*, both human and MLLM supporters of the Plain flashcard emphasized that the focused photograph of a person sitting on the bench made the referent easier to identify and fit naturally with the wedding context in the sample sentence. At the same time, the disagreement cases shown in the bottom three rows of Table 4 reveal important differences in what humans and MLLMs attend to. Human respondents often valued simpler wording, contextual naturalness, and semantic fit. For example, for *to restore*, some preferred the Plain flashcard because its sentence was easier to understand, even when the illustration itself was less clear. For *sombre*, some participants preferred the Plain version because people with serious expressions helped convey the intended nuance. For *a pew*, human comments suggest that comparison-based layouts can sometimes feel unnecessary or even interfere with interpretation. By contrast, MLLM rationales consistently emphasized structural explicitness, visual contrast, and split-panel organization, preferring the Rich flashcard when these design features made the target meaning more overt. These qualitative differences help explain why models tend to overrate the Rich flashcard for *sombre* and *a pew*, whereas human judgments remain more divided.

Main Divergence Points in Human Feedback Analysis of human feedback revealed that the main points of divergence among respondents concerned the presence of people in the image and the use of comparison-based layouts. To investigate these patterns and their implications for possible customization of flashcard-style elaboration generation to individual learner needs, we compare the background profiles of respondents who explicitly referred to either people or comparison-style layouts in their comments. The results shown in Figure 5 suggest that preferences regarding both the inclusion of people and comparison-based visual support are somewhat sensitive to learner profile. Comments favoring people in the image are concentrated among younger respondents and occur slightly more often among beginner-to-intermediate learners, whereas comments disfavoring people show a relatively larger share of older respondents and advanced learners. Comparison-based representations, by

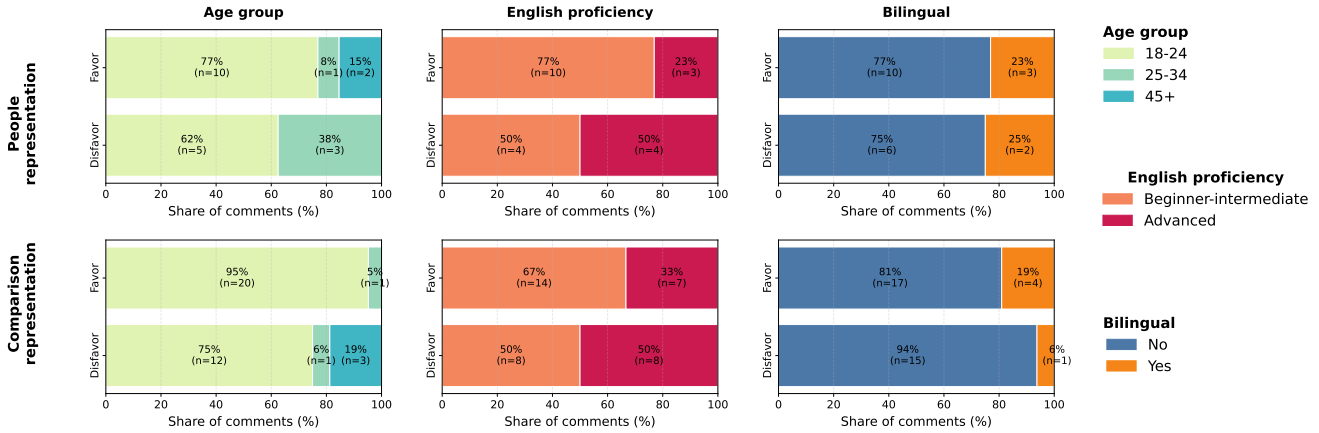

Figure 5. **Background profiles of comments on people in images and comparison-based representations**. The top row shows comments favoring or disfavoring the presence of people; the bottom row shows comments favoring or disfavoring comparison-based visuals. Each subplot gives the distribution by age, English proficiency, and bilingual status, with percentages and raw counts.

contrast, are generally well received among younger participants, with favorable comments clearly dominating in both the age and bilingual-status breakdowns. However, older and more advanced participants appear less likely to favor this design style.

POS	Human comments
verb (to restore)	"Since restore is a verb, I thought it would be easier to understand if I had **an image of a person performing the action**."
adjective (sombre)	"Because "sombre" describes a momentary quality of something, an image like **a snapshot that captures a single moment** is easier to understand."
noun (a pew)	"Because the noun pew **is used together with the verb "sat"**, a card with a photo that makes it immediately clear what object someone is sitting on is easier to understand."

Table 5. Suggested visual representations for different parts of speech based on human comments.

Visual Representations by Part of Speech To investigate which types of visual representation work best for different parts of speech, we collected representative human comments, as shown in Table 5. The results suggest that participants associate different parts of speech with different preferred visual representations. Notably, this type of rationale did not appear in the MLLM comments. For verbs, some comments indicated a preference for images that depict a person performing the action, making the dynamic nature of the verb more immediately interpretable. For adjectives, some participants preferred snapshot-like images that capture a salient moment, mood, or quality, which helps convey descriptive meaning more directly. For nouns, participants tended to prefer images that clearly present the target object together with an action commonly associated with it, as in the example of *a pew*, where sitting helps identify the object. Taken together, these comments suggest that visual comprehensibility depends not only on image quality, but also on how well the representation matches the semantic role of the target keyword.

Modality Preference Across the additional analyses, a consistent pattern emerges: although image support remains the strongest single modality overall, many learners report that image and caption contribute equally, and caption reliance increases as familiarity decreases, as illustrated in Figure 6. More specifically, as familiarity declines, reliance on image-only support generally weakens and preferences shift toward multimodal integration: for *to restore*, the image-dominant pattern among familiar learners (Yes: 67.2%) becomes more balanced in the "Heard" group (Both equally: 44.4%); for *a pew*, multimodal preference is strongest in lower-familiarity groups (Heard: 50.0%; No: 43.3%); and for *sombre*, the "No" group shows increased caption reliance (Mostly caption: 39.7%). These results suggest that combined image+text designs are more robust for unfamiliar vocabulary than image-led designs alone.

6. Comprehensibility Rubric for MES

Based on the qualitative evidence gathered in MES-Bench, we propose a preliminary comprehensibility rubric: a human-perception-centered set of criteria for evaluating the MES task. Table 6 summarizes the proposed evaluation dimensions for both modalities, with the upper blue section covering caption-related criteria and the lower yellow section covering image-related criteria. For elaborative caption evaluation, in line with the qualitative feedback, we argue that evaluation should consider the length of the explanation (*Explanation Brevity/Conciseness*), whether the caption aligns with the image (*Image-text Alignment/Referential Clarity*), whether it includes concrete examples (*Presence of Concrete*

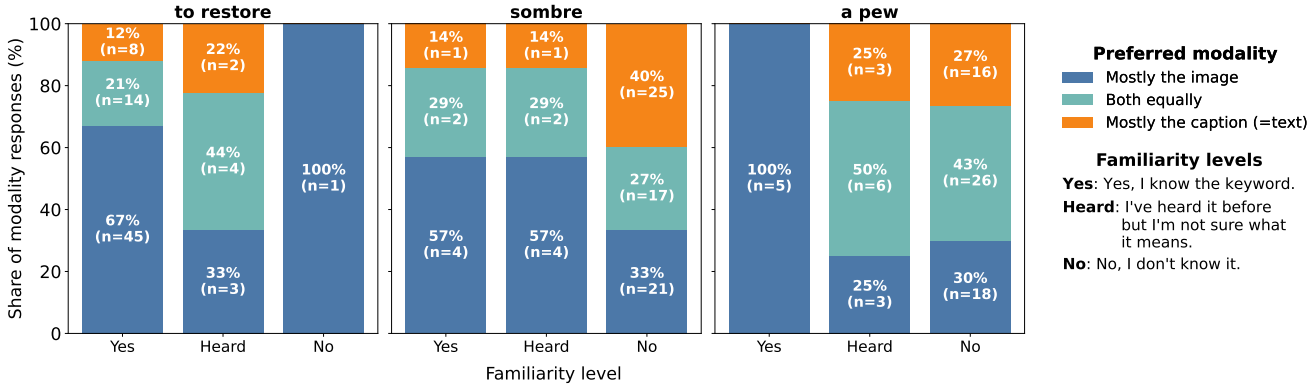

Figure 6. **Modality preference by familiarity across keywords.** Stacked bars for *to restore*, *sombre*, and *a pew* show the distribution of image-led, caption-led, and multimodal (*both equally*) preferences across familiarity levels (*Yes*, *Heard*, *No*); labels indicate percentages and respondent counts.

Examples), useful paraphrases, or lexical cues that support inference of the target meaning (*Use of paraphrasing/lexical support*), and whether the main explanation appears early enough in the caption (*Placement of the explanation*).

On the image side, we conclude that the rubric should consider whether the scene is realistic (*Realism/image quality*), whether the visual focus is placed on the main object or action associated with the keyword (*Focus/Perspective*), whether the image is sufficiently visible and easy to inspect (*Lightning/Visibility*), whether distracting background elements are present (*Level of Detail*), and whether additional objects genuinely help extend the context needed to understand the target keyword (*Presence of Visual Cues*). The rubric should also account for semantic ambiguity, that is, whether the image invites multiple interpretations (*Precise Semantic-fit/Ambiguity*), as well as layout structure and object placement, which may affect how intuitively the meaning can be grasped (*Layout Structure*).

7. Discussion and conclusion

This study introduces MES-Bench, a benchmark for comparing how human learners and MLLMs evaluate the comprehensibility of translation-free, multimodal flashcard-style elaborations for advanced vocabulary. Motivated by systematic gaps in how both groups interpret "comprehensibility," we also propose a human-centered rubric grounded in our findings.

Quantitatively, human–MLLM divergence was most pronounced for flashcards with strong visual contrast or multi-panel layouts. Models consistently favored explicit structural cues (e.g., before/after comparisons, antonym pairings), whereas human ratings were more moderate and variable, suggesting a more nuanced evaluation process. Qualitative analysis further clarifies this gap. While both humans and MLLMs valued clear text-image links, human respondents often preferred simpler, more concise explanations with the main information presented at the very beginning. By contrast, MLLMs tended to favor explicitness and more descriptive explanations. On the image side, participants favored realistic, keyword-focused images with enough detail to support understanding but without distraction. They were also more sensitive than MLLMs to linguistic nuance and visual ambiguity. Preferences further differed by layout: some humans preferred single-frame images, whereas MLLMs almost always favored contrast framing and multi-panel designs.

Human feedback also revealed a clear split around two factors: the presence of people and comparison-style layouts. Learner-profile analysis suggests that these preferences may vary by age and proficiency, with younger beginner-to-intermediate learners tending to favor people and contrastive multi-panel designs, and older, more advanced learners more often preferring object-focused single-frame images. A related pattern emerged when we examined representations by part of speech. Different visual formats appeared to support learners differently: for verbs, some learners preferred depictions of actions in progress, whereas for adjectives and nouns they more often favored static representations. Additional learner-side analysis showed that, although images were the dominant single cue overall, preferences shifted toward combined image-and-caption support as keyword familiarity decreased, indicating that multimodal designs are more robust than image-only designs for less familiar vocabulary.

These findings suggest that that comprehensibility in MES emerges from the interaction of textual and visual support, and that without human-centered criteria, evaluation may favor features that seem explicit to models but do not necessarily help learners. We therefore argue that future MES evaluation should move beyond coarse preference judgments and assess these dimensions explicitly, enabling a more principled analysis of why a flashcard succeeds or fails for learners and opening the door to future customization. An important direction for future work is to extend this analysis to a broader range of vocabulary items.

Evaluation dimension	Supportive human comments
Caption	
Explanation Brevity / Conciseness	"Almost same, but since the sentence on the right card **is shorter**, it might be easier to understand." "In terms of the explanation, the left one **uses less text** and makes it easier to infer the meaning through multiple paraphrases (...)."
Image-text Alignment / Referential Clarity	"the left one is clearer because it because the scene it presents is more **similar to the sample sentence**." "In the explanation for the left one, it was confusing that it uses "they" even though there is only one person in the image..."
Presence of Concrete Examples	"I understood the meaning of "sombre" after reading the **"for example ~"** in the context on the right card." "I don't really need the opposite example."
Placement of the Main Explanation	"It is easy to understand because **what the word refers to is explained first**, and the visual emphasis in the image also helps." "The image is a little hard to understand, but the first sentence gives a precise explanation, so it is easy to understand."
Use of Paraphrasing / Lexical Support	"Because the word "damaged" in the original sentence also makes it easy to infer the nuance of restoring a damaged painting." "The second card seems to make it easier to grasp the meaning of the word by also taking its **antonym** into account."
Image	
Realism / Image Quality	"Because the image on the left is **more realistic**, like a photograph, I felt it was easier to grasp what the sentence was referring to." "The reason is that the photo is very clear and realistic, so I can immediately recognize what a pew looks like in an actual church."
Lighting / Visibility	"pic 1 is clear while pic 2 is **yellow and dim**" "The right side flashcard is **partially obscured by a shadow**, making it difficult to see the whole thing."
Focus / Perspective	"With the image on the left, we wouldn't know the subject is a bench unless there was an explanation, but in the image on the right, **the bench is clearly the focus.**" "Because the picture on the left more clearly shows the long chair within the boundary."
Level of Detail	"So much **irrelevant things in the picture**. The second one is clearer." "In this case I had to read the caption because the images have many information (ex., people, bench, building) and it's hard to tell which one you're pointing out."
Precise Semantic Fit / Ambiguity	"The second image can also be interpreted as coloring a painting, so it is not suitable for conveying the exact meaning of the word." "The second image looks like a contrast between brightness and darkness, so **it is hard to notice without reading the explanation** carefully that only the darker side represents the meaning of the word."
Presence of Visual Cues	"In the right one, **the arrow** made my understanding clearer." "In the left image, the information 'painter' and 'tools' stands out first, so it takes an extra moment to arrive at the idea of 'restoration.'"
Layout Structure	"By showing **the contrasting images**, it was easy to understand what was being referred to." "With the left one, you can't immediately tell what is happening at a glance. The right one is more **diagram-like** and easier to understand."

Table 6. **Proposed comprehensibility rubric for the MES task, grounded in human comments from MES-Bench.** The table summarizes recurring factors that participants attended to when judging whether a flashcard helped them understand the target word. Blue rows denote caption-related evaluation dimensions, while yellow rows denote image-related evaluation dimensions.

References

[1] Anthropic. Claude haiku documentation, 2025.

[2] Anthropic. Claude sonnet documentation, 2025.

[3] Shuai Bai et al. Qwen2.5-vl technical report, 2025.

[4] Shuai Bai et al. Qwen3-vl technical report, 2025.

[5] Sami Baral, Li Lucy, Ryan Knight, Alice Ng, Luca Soldaini, Neil Heffernan, and Kyle Lo. DrawEduMath: Evaluating vision language models with expert-annotated students' hand-drawn math images. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6902–6920. Association for Computational Linguistics, 2025.

[6] Luca Benedetto and Paula Buttery. Towards CEFR-targeted text simplification for question adaptation. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 150–157. INCOMA Ltd., Shoumen, Bulgaria, 2025.

[7] Redacted by arXiv. The llama 4 herd: Architecture, training, evaluation, and deployment notes, 2026.

[8] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

[9] Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327. Association for Computational Linguistics, 2024.

[10] Alison Chi, Li-Kuang Chen, Yi-Chen Chang, Shu-Hui Lee, and Jason S. Chang. Learning to paraphrase sentences to different complexity levels. *Transactions of the Association for Computational Linguistics*, 11:1332–1354, 2023.

[11] Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025.

[12] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163: 90–100, 2017.

[13] David Dinucu-Jianu, Jakub Macina, Nico Daheim, Ido Hakimi, Iryna Gurevych, and Mrinmaya Sachan. From problem-solving to teaching problem-solving: Aligning LLMs with pedagogy using reinforcement learning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 272–292. Association for Computational Linguistics, 2025.

[14] Warwick B. Elley and Francis Mangubhai. The impact of reading on second language learning. *Reading Research Quarterly*, 19(1):53–67, 1983.

[15] Björn Engelmann, Fabian Haak, Christin Katharina Kreutz, Narjes Nikzad Khasmakhi, and Philipp Schaer. Text simplification of scientific texts for non-expert readers, 2023.

[16] Gemma Team et al. Gemma 3 technical report, 2025.

[17] Google. Gemini 3 pro model card, 2025.

[18] Theo Guidroz, Diego Ardila, Jimmy Li, Adam Mansour, Paul Jhun, Nina Gonzalez, Xiang Ji, Mike Sanchez, Sujay Kakarmath, Mathias MJ Bellaiche, Miguel Ángel Garrido, Faruk Ahmed, Divyansh Choudhary, Jay Hartford, Chenwei Xu, Henry Javier Serrano Echeverria, Yifan Wang, Jeff Shaffer, Eric, Cao, Yossi Matias, Avinatan Hassidim, Dale R Webster, Yun Liu, Sho Fujiwara, Peggy Bui, and Quang Duong. Llm-based text simplification and its effect on user comprehension and cognitive load, 2025.

[19] Ching Nam Hang and Sin Man Ho. Personalized vocabulary learning through images: Harnessing multimodal large language models for early childhood education. In *2025 IEEE Integrated STEM Education Conference (ISEC)*, pages 1–7, 2025.

[20] Ching Nam Hang, Chee Wei Tan, and Pei-Duo Yu. Mcqgen: A large language model-driven mcq generator for personalized learning. *IEEE Access*, 12:102261–102273, 2024.

[21] Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[22] Stephen D. Krashen. *Second Language Acquisition and Second Language Learning*. Pergamon Press, 1981.

[23] Stephen D. Krashen. *Principles and Practice in Second Language Acquisition*. Pergamon Press, 1982.

[24] Stephen D. Krashen. *Explorations in Language Acquisition and Use: The Taipei Lectures*. Heinemann, 2003.

[25] Stephen D. Krashen and Tracy D. Terrell. *The Natural Approach: Language Acquisition in the Classroom*. Alemany Press, 1983.

[26] Joon-Young Lee, Sang-Hyun Park, et al. Learnmate: Enhancing online education with llm-powered personalized learning plans and support. *arXiv preprint arXiv:2503.13340*, 2025.

[27] Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11286–11315. Association for Computational Linguistics, 2024.

[28] Jionghao Lin, Eason Chen, Ashish Gurung, and Kenneth R. Koedinger. Mufin: A framework for automating multimodal feedback generation using generative artificial intelligence. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, page 550–552. Association for Computing Machinery, 2024.

[29] Yiqi Liu, Nafise Moosavi, and Chenghua Lin. LLMs as narcissistic evaluators: When ego inflates evaluation scores. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12688–12701. Association for Computational Linguistics, 2024.

[30] Yuhang Liu, Mo Zhang, Zhaoyi Cheng, and Sarah Ebling. Document-level simplification and illustration generation multimodal coherence. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 19–35. Association for Computational Linguistics, 2025.

[31] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: multimodal reasoning via thought chains for science question answering. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2022.

[32] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024.

[33] Beniko Mason. Guided ssr before self-selected reading. *Shitennoji University Bulletin*, 67:445–456, 2019.

[34] Beniko Mason and Nobuyoshi Ae. 70 hours of comprehensible input = 286 hours of traditional instruction. pages 251–259, 2023.

[35] Beniko Mason and Stephen D. Krashen. Extensive reading in english as a foreign language. *System*, 25(1):91–102, 1997.

[36] Beniko Mason and Stephen D. Krashen. Self-selected reading and toeic performance: Evidence from case histories. *Shitennoji University Bulletin*, 63:469–475, 2017.

[37] Ibomoiye Domor Mienye and Theo G. Swart. Chatgpt in education: A review of ethical challenges and approaches to enhancing transparency and privacy. *Procedia Computer Science*, 254:181–190, 2025. International Conference on Digital Sovereignty (ICDS).

[38] Mistral AI. Mistral small 3.1, 2025.

[39] Jalal Nikoopour and Alireza Kazemi. Vocabulary learning through digital flashcards versus paper flashcards: An experimental study. *Journal of Language Teaching and Research*, 5 (6):1364–1372, 2014.

[40] NVIDIA. Nvidia nemotron nano 12b v2 vl model card, 2025.

[41] OpenAI. Dall-e 3 system card, 2023.

[42] OpenAI. Gpt-5.1 instant and gpt-5.1 thinking system card addendum, 2025.

[43] OpenAI. Openai gpt-5 system card, 2025.

[44] OpenAI et al. Gpt-4o system card, 2024.

[45] Belkiss Souayed, Sarah Ebling, and Yingqiang Gao. Template-based text-to-image alignment for language accessibility a study on visualizing text simplifications. In *Proceedings of the Fourth Workshop on Text Simplification, Accessibility and Readability (TSAR 2025)*, pages 1–18. Association for Computational Linguistics, 2025.

[46] Neha Srikanth and Junyi Jessy Li. Elaborative simplification: Content addition and explanation generation in text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5123–5137. Association for Computational Linguistics, 2021.

[47] Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450. Association for Computational Linguistics, 2024.

[48] Yating Wu, William Sheffield, Kyle Mahowald, and Junyi Jessy Li. Elaborative simplification as implicit questions under discussion. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5525–5537. Association for Computational Linguistics, 2023.

[49] Isaac Xodabande and Mohammad Reza Atai. Using digital flashcards for vocabulary learning: Effects on retention and motivation. *Computer Assisted Language Learning*, 33(7): 1–23, 2020.

[50] Songlin Xu, Xinyu Zhang, and Lianhui Qin. Eduagent: Generative student agents in learning, 2024.

[51] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.