Creating vs. Viewing Multimodal Videos: Effects on L2 Vocabulary Retention

Chao Xu (chao.xu@studenti.unipd.it)[1,2] & Yan Gu[3,4]
[1]Department of Linguistic and Literary Studies, University of Padua
[2]Department of Foreign Languages, Chaohu University
[3]Department of Psychology, University of Essex
[4]Department of Experimental Psychology, University College London

Abstract

This study examined whether learner-generated multimodal videos enhance L2 word retention beyond multimodal exposure alone and beyond traditional teacher-led instruction. To isolate production from exposure, 164 Chinese university EFL learners were taught 39 unfamiliar words in one of three conditions: video creation (each learner produced two short videos integrating visual, auditory, linguistic, bodily, and affective resources), video viewing (simply watched the same videos), or traditional instruction (teacher-led explanations and examples). Retention was assessed with immediate and delayed post-tests, with previously known items excluded before instruction. Both video-based conditions outperformed traditional instruction, with the largest gains in the video creation condition. The creation advantage extended beyond self-authored items: Creators also outperformed viewers on words encountered only through subsequent in-class viewing, suggesting that creating videos alters attention to and use of multimodal cues during later viewing. Brief learner-generated multimodal production therefore supports deeper encoding and more durable lexical learning than exposure alone.

Keywords: learner-generated content; multimodal learning; vocabulary retention; EFL; embodied cognition

Introduction

Vocabulary knowledge is a foundational component of L2 ability, underpinning comprehension and supporting fluent language use (Nation, 2013; Schmitt, 2010; Webb & Nation, 2017). In the classroom, vocabulary is typically taught through a mix of intentional approaches (e.g., explicit explanation of word meaning and form and deliberate study with word cards), and incidental or meaning-focused approaches (e.g., learning from contextualized input through reading or listening) (Hulstijn, 2001; Nation, 2013). For example, Nation (2013) distinguishes strands of meaning-focused input, meaning-focused output, language-focused learning, and fluency development, each implying different routes to vocabulary growth. Relatedly, Webb and Nation (2017) discuss how deliberate study (including word-card learning and repeated encounters) and contextual exposure can jointly support lexical development. Despite this methodological diversity, an enduring challenge is ensuring that newly learned form–meaning mappings remain accessible beyond the immediate lesson, rather than reflecting only short-lived performance gains. From a cognitive perspective, Craik and Lockhart's (1972) levels-of-processing framework predicts stronger memory when encoding is more elaborative, and Laufer and Hulstijn's (2001) involvement load hypothesis similarly argues that tasks inducing greater "need, search, and evaluation" should yield stronger vocabulary learning. In practice, instructional activities that require learners to elaborate a word's meaning and evaluate how cues map onto that meaning are likely to induce the kind of deep processing and high involvement these frameworks emphasize. Multimodal vocabulary instruction is one practical way to create these conditions, enriching verbal explanation with semantically aligned imagery, sound, gesture/action, and affective framing.

From the input perspective, multimodal vocabulary instruction (including short explanatory videos) can be seen as an attempt to implement the verbal-nonverbal pairing, which can provide learners with multiple retrieval routes and support their later access to new lexical knowledge (e.g., dual coding theory, Paivio, 1991). Extending this rationale, Mayer's cognitive theory of multimedia learning specifies when such benefits are most likely to occur: Learning gains depend on the coherent coordination of modalities rather than the mere addition of media (Mayer, 2009). In particular, Mayer and Moreno (2003) highlight that poorly aligned multimedia can impose unnecessary cognitive load (Sweller, 1988), whereas well-designed materials support learning by helping learners select relevant information, organize it into coherent representations, and integrate it with prior knowledge (Mayer, 2009). Consistent with these principles, early L2 work on multimedia glosses/annotations provided evidence that combining verbal annotations with visual support (pictures and/or video) can facilitate L2 vocabulary learning (Chun & Plass, 1996; Plass et al., 1998). More recently, a meta-analytic synthesis incorporating document co-citation analysis reported an overall large effect of multimedia glosses on L2 vocabulary acquisition with effect sizes varying by learner proficiency and test tasks (e.g., larger for recognition tests than for vocabulary recall tests) (Mohsen et al., 2024). A related line of evidence comes from audiovisual input, where captions have been shown to support recognition of written word forms and the learning of word meanings (Sydorenko, 2010), and meta-analytic findings indicate that captioned viewing yields a reliable medium benefit for incidental L2 vocabulary learning (Kurokawa et al., 2025).

Despite growing evidence that well-designed multimodal input can support L2 vocabulary learning, most multimedia approaches treat learners primarily as recipients of explanations rather than as generators of them. A key question, therefore, is whether asking learners to create multimodal explanations yields benefits beyond viewing the

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

same materials. Generative learning theory proposes that learning is enhanced when learners actively generate connections and integrate new information with prior knowledge, emphasizing that these benefits arise from selecting, organizing, and integrating processes (Fiorella & Mayer, 2015; Wittrock, 1974). Applied to vocabulary video creation, this perspective suggests that planning and producing an explanation is likely to require learners to articulate a precise form–meaning mapping and to make informed choices about which cues (e.g., images, narration, contextual scenes, gestures) best convey meaning to an audience, thereby promoting semantic elaboration and stronger memory traces. Relatedly, constructive activities requiring learners to produce outputs that add to or beyond the information should produce deeper learning than active or passive activities such as watching, because they more reliably induce inferential and integrative processing (Chi & Wylie, 2014). This is consistent with classic memory research demonstrating that self-generation can enhance later recall compared with reading the same information, a phenomenon known as the generation effect; generating material oneself is thought to encourage more active elaboration and stronger retrieval cues than passive exposure (Slamecka & Graf, 1978).

Additionally, vocabulary video creation may also engage learners' sensorimotor and affective systems through gestures, facial expressions, and enactments, which could strengthen lexical encoding. In line with embodied and grounded views of cognition, word meanings are partly supported by perceptual, motor, and affective systems rather than being represented only as amodal symbols (Barsalou, 2008; Meteyard et al., 2012). Within this tradition, gesture is not treated as a decorative add-on but as a mechanism that links language and cognition in systematic ways (Kita & Emmorey, 2023), and sensorimotor involvement can be a principled design dimension in learning and instruction (Castro-Alonso et al., 2024). For example, embodied approaches are found to support language learning across contexts and tasks (Jusslin et al., 2022), and there is an enactment advantage, with action at encoding yielding better later memory than verbal learning or observation (Roberts et al., 2022). In L2 vocabulary learning, using meaningful gesture or enactment also has learning benefits (Kelly et al., 2009; Macedonia & Knösche, 2011; Zhou & Gu, 2024), and such reliable advantages are both present in enacting and observing gestures during L2 vocabulary learning (Oppici et al., 2023; see also Macedonia, 2025). These claims map naturally onto learner-generated vocabulary videos, which often rely on gestures, facial expressions, and action scenes, and may therefore strengthen form–meaning mapping by binding lexical items to sensorimotor and affective cues (Barsalou, 2008; Kita & Emmorey, 2023).

Indeed, recent classroom studies suggest that learner-generated multimodal projects can support vocabulary learning. For example, there are vocabulary benefits from students' self-made digital stories (Nami & Asadnia, 2024), and learners who created vocabulary videos had greater gains than peers who only used the videos in a flipped EFL setting (Bobkina et al., 2025).

However, isolating the creation advantage is methodologically challenging. First, creation typically requires additional out-of-class work. Engaged learning time is a well-established determinant of achievement (Carroll, 1963; Fisher et al., 1979; Karweit, 1984), which makes time-on-task differences a plausible alternative explanation. This concern is especially relevant for vocabulary, because repeated encounters and sustained engagement are central to lexical development (Nation, 2013). Second, creation and viewing conditions can differ in learners' exposure to target items (e.g., how often and in what input distributions they are encountered), and such frequency/distributional differences can affect learning and retention (Ellis, 2002). Third, creation effects may be inflated by item-specific authoring benefits, given well-established memory advantages for self-generated and overtly produced material (MacLeod et al., 2010; Slamecka & Graf, 1978).

Furthermore, it is unknown whether the constructive activity of creating a video may also support broader learning processes. Jeannerod (2001) argues that action observation recruits internal motor simulation, and prior motor expertise can modulate responses to observed actions (Calvo-Merino et al., 2005). If so, engaging in creation may shift learners into a more generative stance, and thus process peers' videos more deeply, even for words they did not produce any videos.

In short, designs should hold the in-class input constant across multimodal conditions and distinguish words that learners authored from words they only encountered via peers' videos in the classroom lecture. To address these issues, we compared learning outcomes across three conditions: video creation, video viewing, and traditional instruction (teacher-led explanation of word meaning with example sentences, without video materials), while matching in-class exposure time. Crucially, we tracked whether each word was self-produced or view-only to test whether any creation advantage extends beyond authored words and persists over time. We asked whether (1) video-based multimodal instruction outperforms traditional instruction on immediate and delayed tests; (2) when in-class input is matched, video creation yields an advantage over viewing; and (3) forgetting from immediate to delayed testing differs across conditions. Because multimodal materials can provide multiple retrieval cues, we predicted that both video conditions would yield higher immediate retention than traditional instruction and would reduce forgetting over time. Additionally, if producing changes how learners attend to and use multimodal cues during subsequent viewing, creators should also outperform viewers on words encountered only through later in-class viewing.

Method

Participants

164 second-year university EFL learners in China participated in the study (M = 19.38 years; 57 men, 107 women). Following a quasi-experimental design (Shadish et al., 2002), participants were recruited from three intact classes and assigned by class to one of three instructional conditions: video creation (n = 71), video viewing (n = 49), or traditional instruction (n = 44). All participants provided informed consent. Ethical approval was obtained from the University of Padua.

Materials

Target vocabulary The instructional materials consisted of 39 English target words at the B2–C1 level that were expected to be unfamiliar to the participants. These words were selected from an initial pool of 100 items through a pilot familiarity check administered to 24 high-proficiency students from a comparable EFL learner population who did not participate in the main study. Words endorsed as familiar by at least 30% of pilot participants were excluded, resulting in the final set of 39 target words. In the main study, any items reported as known before class were further excluded from the analyses.

Student-generated videos In the two video-based conditions, the instructional materials were student-generated vocabulary videos. In the video-creation condition, learners made two one-minute videos, each explaining one target word with multimodal resources such as images, text, narration, gestures, and contextualized scenes. The instructor then selected one representative video for each target word based on semantic accuracy, clarity, and audiovisual quality. The resulting 39 videos were used as standardized materials for the video-viewing group and for the in-class viewing phase of the video-creation group. Figure 1 shows an example frame from a student-generated vocabulary video for the target word *derive*.

Figure 1: Example frame from a student-generated vocabulary video for *derive*, illustrating multimodal form–meaning mapping through on-screen text, expressive enactment, and contextual audiovisual cues (background music and situational setting).

Procedure

Prior to the classroom vocabulary lesson, all participants completed baseline measures of English proficiency and working memory in class (see details in Measures). The same 39 target words were taught in all conditions. In the video-creation condition, each learner was assigned two target words and produced two short videos prior to the classroom lesson, one for each assigned word. To reduce potential cross-item contamination during the pre-class production period (e.g., peer discussion or early access to others' materials), students in the creation condition were explicitly instructed not to share their assigned target words, scripts, or videos with peers before the in-class session. After collecting the submissions, the instructor selected one representative video for each target word to assemble a standardized set of 39 classroom videos. In the video-viewing condition, learners did not create videos but watched this same set of 39 videos in class. In the traditional-instruction condition, the instructor taught the same target words through pronunciation, explanation, and example sentences, with one minute allocated to each target word. Participants completed an immediate post-test after instruction and a delayed post-test five days later.

Measures

English proficiency Overall English proficiency was assessed using the Oxford Quick Placement Test (UCLES, 2001) and entered as a covariate in the analyses. A one-way ANOVA indicated no significant between-group differences in baseline proficiency across the video creation (M = 29.5, SD = 8.64), video viewing (M = 30.4, SD = 9.56), and traditional instruction groups (M = 32.1, SD = 5.03), $F(2, 161) = 1.34$, $p = .26$.

Working memory Working memory was assessed with a composite test adapted from an online assessment Arealme website (https://www.arealme.com/memory-test/cn/), comprising 13 items (9 verbal, 3 visual, 1 numerical). Scores were summed and entered as a covariate in the analyses. A one-way ANOVA indicated no significant between-group differences in working memory across the video creation (M = 89, SD = 9.58), video viewing (M = 91.2, SD = 9.71), and traditional instruction groups (M = 91.6, SD = 9.63), $F(2, 161) = 1.25$, $p = .29$.

Vocabulary retention Vocabulary learning was assessed with an immediate test and a 5-day delayed post-test. Both tests used multiple-choice items targeting form–meaning mapping for the 39 target words. To reduce predictability, each test also included four untaught filler items, which were not included in the main analyses. Each item included a familiarity check to identify pre-existing knowledge; items marked as known were excluded. Learners in the video creation condition also indicated whether they had

personally created a video for each target word, enabling analyses of self-produced (words for which a learner created a video) versus non-produced items (words for which a learner did not create a video, but would learn by viewing peers' videos in the classroom lecture).

Data Analysis

Vocabulary-test responses were analyzed at the item level (correct = 1, incorrect = 0) using generalized linear mixed-effects models (GLMMs) with a logit link (Jaeger, 2008) in R (R Core Team, 2025) using the lme4 package (Bates et al., 2015), with analyses conducted in the RStudio IDE (Posit Team, 2025). The primary model included instructional condition (video creation, video viewing, traditional instruction), test time (immediate vs. delayed), and their interaction as fixed effects to evaluate overall learning and differential forgetting across conditions. Learners' English proficiency scores (Oxford_z) and working-memory scores (Memory_z) were standardized (z-scored) and included as covariates to control for individual differences.

The models included random intercepts for participants and word items. More complex random-effects structures with by-participant slopes for test time failed to converge. Therefore, the final models included random intercepts for participants and word items only. Items marked as known (M = 7.18, SD = 7.51) prior to instruction (but not due to the creation of videos for the video-creation condition) were excluded from all analyses. After excluding these known items, participants contributed an average of 31.8 analyzable items (SD = 7.51). In addition to the main model, we conducted a planned supplementary between-condition analysis focusing on self-produced items to examine the item-specific benefit of authoring. To test whether any creation advantage generalized beyond authored words, we then repeated the between-condition comparisons after excluding self-produced items from the video-creation group dataset, focusing on non-produced items shared across conditions.

Results

Primary analyses focus on non-produced items shared across conditions (known items excluded; self-produced items removed from the creation group). Table 1 reports participant-level descriptive statistics (n, M, SD) for accuracy on non-produced target items at the immediate and Day 5 tests.

Results from the combined binomial GLMM (10,152 trial-level observations; Table 2) showed effects of instructional condition, test time, and their condition-by-time interaction. At the immediate test, both video conditions outperformed traditional instruction. Accuracy declined from immediate to Day 5 in the traditional condition (OR = 0.32, 95% CI [0.27, 0.39], $p < .001$), but this decline was attenuated in the viewing condition (Viewing × Day 5: OR = 1.34, 95% CI [1.04, 1.72], $p = .022$) and further attenuated in the creation condition (Creation × Day 5: OR = 1.54, 95% CI [1.21, 1.96], $p < .001$). Planned contrasts showed that creation outperformed viewing at both time points (immediate: OR = 2.13, 95% CI [1.45, 3.13], $p < .001$; Day 5: OR = 2.44, 95% CI [1.68, 3.56], $p < .001$). Figure 2 visualizes the fixed-effect odds ratios reported in Table 2.

Table 1: Mean and SD (bracket) of test outcomes for non-produced target items.

Statistic	Creation	Viewing	Traditional
Immediate test %	82.6 (14.6)	73.2 (18.2)	50.9 (26.2)
Day 5 test %	73.8 (13.5)	59.2 (15.0)	30.9 (17.5)
Known items excluded	6.2 (6.8)	10.1 (9.1)	5.6 (5.7)
Remaining target items	30.8 (6.8)	28.9 (9.1)	33.4 (5.7)

Note: known items were excluded; self-produced items were removed from the creation group.

Table 2: Fixed-effect estimates from the combined binomial GLMM on non-produced items (known items excluded; self-produced items removed from the creation group).

Predictor	OR [95% CI]	*p*
Viewing (vs. Traditional)	3.60 [2.37, 5.47]	< .001
Creation (vs. Traditional)	7.66 [5.17, 11.37]	< .001
Day 5 (vs. Immediate)	0.32 [0.27, 0.39]	< .001
Viewing × Day 5	1.34 [1.04, 1.72]	.022
Creation × Day 5	1.54 [1.21, 1.96]	< .001
Lang Prof (Oxford z-score)	1.13 [0.97, 1.32]	.111
Working Memory (z-score)	1.16 [0.99, 1.35]	.060

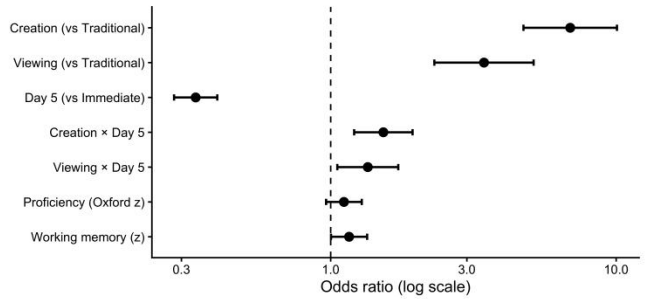

Figure 2: Fixed effects from the binomial GLMM predicting vocabulary test accuracy. Points indicate odds ratios and horizontal bars indicate 95% confidence intervals (Wald). The vertical dashed line marks OR = 1. Traditional and Immediate are reference levels.

We then conducted a planned supplementary between-condition analysis focusing on authored words to examine the item-specific benefit of authoring. As expected, learners in the video-creation condition showed near-ceiling accuracy for self-produced items on the immediate test (M = 97.9%, SD = 10.1%), substantially higher than the viewing group (M = 73.2%, SD = 18.2%, OR = 29.77, $p < .001$) and the traditional instruction group (M = 50.9%, SD = 26.2%, OR = 114.59, $p < .001$) on the same items. The same pattern held at the delayed post-test (97.2% vs. 59.2% vs. 30.9%),

indicating a strong benefit of creating videos for memorizing the words learners authored.

Crucially, the creation advantage was not limited to self-produced items. To test generalization beyond authored words, we repeated the between-condition comparisons after excluding self-produced items from the creation group, focusing on non-produced items only (known items excluded). On the immediate test, the creation group still showed the highest accuracy (M = 82.6%, SD = 14.6%), followed by the viewing group (M = 73.2%, SD = 18.2%) and the traditional instruction group (M = 50.9%, SD = 26.2%). The same ordering was observed on the 5-day delayed post-test (creation: M = 73.8%, SD = 13.5%; viewing: M = 59.2%, SD = 15.0%; traditional: M = 30.9%, SD = 17.5%). As shown in Figure 3, model-predicted accuracy declined from immediate to delayed testing in all conditions, but the decline was smaller in the two video conditions, especially in the creation condition.

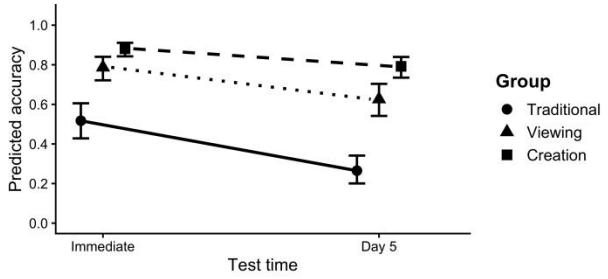

Figure 3: Model-predicted accuracy (± 95% CI) for non-produced items on the immediate and 5-day delayed post-tests across conditions (known items excluded; self-produced items removed from the creation group).

Discussion

Our study asked whether brief learner-generated multimodal vocabulary videos enhance L2 form–meaning retention beyond viewing the same videos and beyond traditional teacher-led instruction, and whether forgetting differs across conditions. The results provide a clear pattern. First, both video-based conditions outperformed traditional instruction on immediate and delayed tests, indicating a robust multimodal advantage. Second, creation outperformed viewing, and this advantage remained when analyses were restricted to non-produced items, suggesting that the benefit of creation extends beyond item-specific authoring. Third, forgetting from the immediate to the delayed test was attenuated in the two video conditions compared with the traditional instruction condition, and was smallest in the creation condition, suggesting more durable learning.

The robust advantage of both video conditions over traditional instruction is consistent with Mayer's (2009) Cognitive Theory of Multimedia Learning (CTML). CTML assumes separate verbal and pictorial channels and predicts better learning when the two are coordinated and integrated. In the traditional condition, learners relied primarily on the teacher's spoken explanations and example sentences, which likely provided fewer systematically coordinated cues for establishing a stable form–meaning mapping. In contrast, the video conditions embedded each word within a brief, concrete scenario where core information was highlighted through on-screen text, auditory narration, and semantically diagnostic imagery. Such multimodal cues can highlight the essential form–meaning mapping and guide learners' attention toward core information, thereby reducing extraneous processing (Fiorella & Mayer, 2021). Relatedly, such coordinated cues support dual coding of the new mapping and strengthen later access (Paivio, 1991). This theoretical rationale finds robust empirical support in a comprehensive meta-analysis of the L2 audiovisual literature (Reynolds et al., 2022). Their systematic synthesis of diverse instructional contexts demonstrated that viewing captioned videos reliably facilitates vocabulary acquisition, with the most pronounced effect sizes observed for intralingual captions and animations. In our study, the on-screen text and dynamic, contextualized enactments likely served a similar cueing function, facilitating the efficient integration of multimodal input. Taken together, the video-based instruction likely created more robust and retrievable "form–meaning episodes" than teacher-led explanations, providing a principled explanation for both the higher accuracy and the attenuated forgetting observed over time.

The especially strong retention observed for self-produced items may also be considered in relation to the production effect. Research on the production effect suggests that material produced aloud or otherwise overtly articulated during study is often remembered better than material processed silently, likely because production adds distinctive encoding cues to the memory trace (MacLeod et al., 2010; MacLeod & Bodner, 2017; Ozubko & MacLeod, 2010). In the present study, learners in the video-creation condition did not merely view multimodal explanations but also overtly produced word-focused explanations for an audience. Although the present design does not disentangle overt production from other components of video creation, the production effect provides a useful complementary framework for interpreting the especially strong retention observed for authored words.

The critical contribution of the present design is that it provides evidence for a creation advantage that generalizes beyond authored words. Even when analyses were restricted to non-produced items learned only through in-class viewing, creators still outperformed viewers. This pattern underscores the added value of generating multimodal explanations through video creation rather than merely receiving them. Crucially, because the in-class stimulus set was identical for creators and viewers, the remaining advantage cannot be attributed to differences in classroom exposure and cannot be explained simply by better memory for authored words. Instead, this production experience may have altered how learners processed multimodal cues during

subsequent viewing. We offer two complementary accounts of this generalized creation advantage.

One plausible explanation is that engaging in multimodal production fostered a more generative mode of processing during subsequent classroom learning. Generative learning theory conceptualizes learning as an active process in which learners select relevant information, organize it into coherent structures, and integrate it with prior knowledge, rather than passively recording input (Wittrock, 1974; Fiorella & Mayer, 2015). In the present task, creating a vocabulary video likely required learners to articulate an explicit form–meaning mapping and to make design decisions about which cues (e.g., imagery, narration, bilingual prompts, gestures) would most effectively convey meaning, while anticipating how an audience would interpret these multimodal cues. Such planning and explanation demands may have promoted deeper semantic elaboration during production and may have carried over to the in-class viewing phase, leading creators to be more engaged in class and to attend more strategically to how peers' videos used multimodal cues to instantiate form–meaning links, even for words they did not produce.

A complementary explanation draws on embodied/grounded cognition, which treats lexical meaning as partly supported by sensorimotor and affective systems (Barsalou, 2008). Converging evidence from the enactment effect literature indicates that performing actions during encoding yields reliable memory benefits relative to viewing or reading alone (Roberts et al., 2022). In our setting, many student videos incorporated enactment-like cues (e.g., gestures, facial expressions, and action-based scenes). For creators, producing such enactment-oriented explanations may have changed how they subsequently processed peers' videos. As Tellier (2008) observed in young L2 learners, the active reproduction of gestures significantly bolsters lexical memorization by functioning as both a visual and a motor modality, which ultimately leaves a richer and more distinctive multimodal trace in memory. Furthermore, having practiced mapping gestures, imagery, and narration onto precise meanings, creators may have become more likely to internally simulate the depicted actions (Jeannerod, 2001; Rizzolatti & Craighero, 2004), a tendency that is often strengthened by prior motor experience (Calvo-Merino et al., 2005). This interpretation is compatible with evidence that overt gestures can support internal computation and that such computation can become internalized with practice. In particular, Chu and Kita (2011) demonstrated that gesture-supported computation in spatial transformation tasks becomes internalized over repeated trials, and that the resulting benefit can persist and generalize even when later gesturing is prohibited. By analogy, creation may have strengthened learners' internal simulation routines during subsequent viewing, thereby supporting encoding and retention even for non-produced items. Prior L2 work likewise suggests that pairing new words with meaning-congruent (iconic) gestures, and having learners enact those gestures during encoding, can improve later retention and accessibility (Kelly et al., 2009; Macedonia & Knösche, 2011). Although we did not directly measure motor engagement or attention, future work could test this mechanism by combining production-time logs with process measures.

Pedagogically, the results highlight a feasible strategy for vocabulary instruction: Assigning learners a small number of target words to explain through short videos, while using a curated set of peer videos for shared viewing, can yield gains beyond viewing alone. The findings also suggest a design implication for student-made vocabulary videos: Combining visual, auditory, linguistic, bodily, and affective elements to convey a target meaning within a concise, memorable episode may help learners build a richer representation and support later retrieval. At the same time, these implications should be interpreted with several caveats. First, the study was conducted with intact classes and class-level assignment rather than individual randomization. Although the groups were comparable on baseline English proficiency and working memory, class-level differences cannot be fully ruled out. Second, the out-of-class time required for video creation was not formally quantified. Third, the quality and coordination of multimodal cues in student-generated videos may vary, and future research should examine whether these factors modulate the strength of the learning gains. Finally, social familiarity with peer-generated videos may have differentially boosted attention in the creation condition. This factor was not measured and could be addressed in future work by using cross-class video sets.

Conclusion

The present study provides clear evidence that learner-generated multimodal videos yield durable benefits for L2 vocabulary learning. Compared with viewing the same videos and with teacher-led instruction, video creation produced the strongest retention and attenuated forgetting over five days. Crucially, the creation advantage extended beyond authored words, with creators also outperforming viewers on non-produced items learned only through in-class viewing. This pattern suggests that creating videos may shape how learners engage with multimodal cues during subsequent viewing, in ways consistent with more generative and grounded learning. While out-of-class effort remains an important consideration, the findings highlight short-form video production by learners as a feasible option for strengthening form–meaning retention in classroom settings.

Acknowledgments

Chao Xu gratefully acknowledges financial support from the China Scholarship Council (CSC) for his doctoral studies. We thank the participating students for their time and cooperation.

References

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bobkina, J., Baluyan, A., & Domínguez Romero, E. (2025). Learning vocabulary by creating: Effects of learner-generated videos on EFL vocabulary acquisition in a flipped classroom. *Education Sciences, 15*(4), 450. https://doi.org/10.3390/educsci15040450

Calvo-Merino, B., Glaser, D. E., Grèzes, J., Passingham, R. E., & Haggard, P. (2005). Action observation and acquired motor skills: An fMRI study with expert dancers. *Cerebral Cortex, 15*(8), 1243–1249. https://doi.org/10.1093/cercor/bhi007

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*(8), 723–733. https://doi.org/10.1177/016146816306400801

Castro-Alonso, J. C., Ayres, P., Zhang, S., de Koning, B. B., & Paas, F. (2024). Research avenues supporting embodied cognition in learning and instruction. *Educational Psychology Review, 36*, Article 10. https://doi.org/10.1007/s10648-024-09847-4

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist, 49*(4), 219–243. https://doi.org/10.1080/00461520.2014.965823

Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General, 140*(1), 102–116. https://doi.org/10.1037/a0021790

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal, 80*(2), 183–198. https://doi.org/10.1111/j.1540-4781.1996.tb01159.x

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Fiorella, L., & Mayer, R. E. (2015). *Learning as a generative activity: Eight learning strategies that promote understanding*. Cambridge University Press. https://doi.org/10.1017/CBO9781107707085

Fiorella, L., & Mayer, R. E. (2021). Principles for reducing extraneous processing in multimedia learning: Coherence, signaling, redundancy, spatial contiguity, and temporal contiguity principles. In R. E. Mayer & L. Fiorella (Eds.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 185–198). Cambridge University Press. https://doi.org/10.1017/9781108894333.019

Fisher, C., Marliave, R., & Filby, N. (1979). Improving teaching by increasing "academic learning time". *Educational Leadership, 37*(1), 52–54.

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 258–286). Cambridge University Press. https://doi.org/10.1017/CBO9781139524780.011

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59*(4), 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Jeannerod, M. (2001). Neural simulation of action: A unifying mechanism for motor cognition. *NeuroImage, 14*(1 Pt 2), S103–S109. https://doi.org/10.1006/nimg.2001.0832

Jusslin, S., Korpinen, K., Lilja, N., Martin, R., Lehtinen-Schnabel, J., & Anttila, E. (2022). Embodied learning and teaching approaches in language education: A mixed studies review. *Educational Research Review, 37*, 100480. https://doi.org/10.1016/j.edurev.2022.100480

Karweit, N. (1984). Time-on-task reconsidered: Synthesis of research on time and learning. *Educational Leadership, 41*(8), 32–35.

Kelly, S. D., McDevitt, T., & Esch, M. (2009). Brief training with co-speech gesture lends a hand to word learning in a foreign language. *Language and Cognitive Processes, 24*, 313–334. https://doi.org/10.1080/01690960802365567

Kita, S., & Emmorey, K. (2023). Gesture links language and cognition for spoken and signed languages. *Nature Reviews Psychology, 2*, 407–420. https://doi.org/10.1038/s44159-023-00186-9

Kurokawa, S., Hein, A. M., & Uchihara, T. (2025). Incidental vocabulary acquisition through captioned viewing: A meta-analysis. *Language Learning, 75*(4), 939–987. https://doi.org/10.1111/lang.12697

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22*(1), 1–26. https://doi.org/10.1093/applin/22.1.1

Macedonia, M. (2025). Your body as a tool to learn second language vocabulary. *Behavioral Sciences, 15*(8), 997. https://doi.org/10.3390/bs15080997

Macedonia, M., & Knösche, T. R. (2011). Body in mind: How gestures empower foreign language learning. *Mind, Brain, and Education, 5*(4), 196–211. https://doi.org/10.1111/j.1751-228X.2011.01129.x

MacLeod, C. M., & Bodner, G. E. (2017). The production effect in memory. *Current Directions in Psychological Science, 26*(4), 390–395. https://doi.org/10.1177/0963721417691356

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect:

Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(3), 671–685. https://doi.org/10.1037/a0018785

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist, 38*(1), 43–52. https://doi.org/10.1207/S15326985EP3801_6

Meteyard, L., Cuadrado, S. R., Bahrami, B., & Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex, 48*(7), 788–804. https://doi.org/10.1016/j.cortex.2010.11.002

Mohsen, M. A., Mahdi, H. S., & Alkhammash, R. (2024). Multimedia glosses and their impact on second language vocabulary acquisition: Insights from a meta-analysis and document co-citation analysis. *Innovation in Language Learning and Teaching, 18*(2), 109–124. https://doi.org/10.1080/17501229.2023.2236084

Nami, F., & Asadnia, F. (2024). Exploring the effect of EFL students' self-made digital stories on their vocabulary learning. *System, 120*, 103205. https://doi.org/10.1016/j.system.2023.103205

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

Oppici, L., Mathias, B., Narciss, S., & Proske, A. (2023). Benefits of enacting and observing gestures on foreign language vocabulary learning: A systematic review and meta-analysis. *Behavioral Sciences, 13*(11), 920. https://doi.org/10.3390/bs13110920

Ozubko, J. D., & MacLeod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1543–1547. https://doi.org/10.1037/a0021015

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie, 45*(3), 255–287. https://doi.org/10.1037/h0084295

Plass, J. L., Chun, D. M., Mayer, R. E., & Leutner, D. (1998). Supporting visual and verbal learning preferences in a second-language multimedia learning environment. *Journal of Educational Psychology, 90*(1), 25–36. https://doi.org/10.1037/0022-0663.90.1.25

Posit Team. (2025). *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA.

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reynolds, B. L., Cui, Y., Kao, C.-W., & Thomas, N. (2022). Vocabulary acquisition through viewing captioned and subtitled video: A scoping review and meta-analysis. *Systems, 10*(5), 133. https://doi.org/10.3390/systems10050133

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27,* 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230

Roberts, B. R. T., MacLeod, C. M., & Fernandes, M. A. (2022). The enactment effect: A systematic review and meta-analysis of behavioral, neuroimaging, and patient studies. *Psychological Bulletin, 148*(5–6), 397–434. https://doi.org/10.1037/bul0000360

Schmitt, N. (2010). *Researching Vocabulary: A Vocabulary Research Manual*. Palgrave Macmillan. https://doi.org/10.1057/9780230293977

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/S15516709COG1202_4

Sydorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning & Technology, 14*(2), 50–73. https://doi.org/10.64152/10125/44214

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture, 8*(2), 219–235. https://doi.org/10.1075/gest.8.2.06tel

University of Cambridge Local Examinations Syndicate. (2001). *Quick Placement Test: Paper and pen test.* Oxford University Press.

Webb, S., & Nation, P. (2017). *How Vocabulary Is Learned.* Oxford University Press.

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist, 11*(2), 87–95. https://doi.org/10.1080/00461527409529115

Zhou, J., & Gu, Y. (2024). The impact of teachers' multimodal cues on students' L2 vocabulary learning in naturalistic classroom teaching. *Proceedings of the Annual Meeting of the Cognitive Science Society, 46*(0). https://escholarship.org/uc/item/8jg192cm