

Article

Beyond the Screen: A Conceptual Framework for Spatial Experiential Vocabulary Learning in Digital Language Education

Hyeree Kang *, Sangsu Choi and Jungyub Woo

IGI Korea, Seoul 08390, Republic of Korea; sangsu.choi@ontwins.com (S.C.); jungyub.woo@ontwins.com (J.W.)

* Correspondence: heidi.kang@ontwins.com

Abstract

Despite the rapid growth of AI-based digital language learning platforms, prevailing approaches to vocabulary instruction have remained relatively unchanged. Flashcards, translation drills, and isolated quizzes continue to dominate, and while they may support short-term retention, they tend to strip away the contextual, physical, and spatial dimensions that research identifies as important for durable language acquisition. This paper proposes spatial experiential vocabulary learning as a pedagogically grounded alternative. Drawing on situated learning theory, embodied cognition, and experiential learning theory, it argues that lasting vocabulary acquisition may depend not only on repeated exposure but also on learners engaging actively with language in meaningful, spatially situated contexts. The central theoretical contribution is the EARR framework, a four-phase learning cycle comprising Encounter, Act, Respond, and Reuse, in which vocabulary is embedded within object-rich spatial environments that invite purposeful interaction rather than passive recognition. The paper develops this framework theoretically and draws out its implications for platform design. It concludes by identifying priorities for future empirical research and by considering how digital language learning environments might be redesigned to better support the conditions under which vocabulary acquisition appears to occur.

Keywords: vocabulary acquisition; spatial learning; embodied cognition; situated learning; experiential learning; digital language learning; EARR framework; object-based learning; immersive environments

1. Introduction

Language learning has become one of the most commercially active areas of educational technology. Platforms offering AI-driven instruction, adaptive feedback, and gamified practice now reach hundreds of millions of learners worldwide. What the research literature has been slower to examine is whether the vocabulary instruction methods underlying these platforms are genuinely suited to how words are learned and whether they are well matched to what vocabulary acquisition actually requires. For many learners, particularly those without access to formal instruction or sustained immersive exposure, a digital platform may be the primary means through which they engage with a second language. That makes what these platforms do, and what they do not do, a matter of some consequence.

The technological sophistication of these platforms has not been matched by comparable progress in their instructional approaches. Vocabulary is widely regarded as foundational to language proficiency, yet the methods dominating digital instruction today

Academic Editors: Yu-Tung Kuo
and Yu-Chun Kuo

Received: 4 June 2026

Revised: 29 June 2026

Accepted: 20 July 2026

Published: 22 July 2026

Copyright: © 2026 by the authors.
Licensee MDPI, Basel, Switzerland.
This article is an open access article
distributed under the terms and
conditions of the [Creative Commons
Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

bear a strong resemblance to those that predated the smartphone. Flashcards, translation drills, and recognition-based quizzes remain the default, not because they are well-supported by the acquisition literature but because they are scalable, measurable, and easy to gamify (Godwin-Jones, 2014; Lan, 2020). These methods may generate short-term retention gains that look encouraging in product analytics, but there is good reason to question whether they produce the kind of word knowledge that allows learners to actually use a language (Nation, 2022; Webb, 2008).

Nation's (2022) account of vocabulary acquisition is useful here. He argued that durable word knowledge develops through repeated encounters across varied contexts, encounters that build up a rich and flexible understanding of how a word behaves rather than a simple form-meaning link. Acquisition, in this view, is a gradual process that depends on contextual variety, active engagement, and meaningful use. Current digital methods tend to provide little of this. They present words in isolation, demand recognition rather than production, and concentrate practice into massed sessions that may work against the very memory mechanisms that support long-term retention.

This paper takes those limitations as its starting point. It proposes spatial experiential vocabulary learning as an alternative and develops a conceptual framework for how such learning might be designed. The argument draws on situated learning theory, embodied cognition, and experiential learning research. Together, these traditions suggest that vocabulary acquisition may depend on learners engaging actively with language within spatially structured, object-rich environments rather than passively encountering words on a screen. The framework proposed here, called EARR, organizes this kind of learning across four phases: Encounter, Act, Respond, and Reuse.

The paper addresses the following question: how might spatial, experiential environments be designed to support vocabulary acquisition in digital language learning contexts? Section 2 reviews current vocabulary instruction methods on digital platforms and identifies their principal limitations. Section 3 develops the theoretical foundations of the proposed approach across four bodies of research. Section 4 introduces the EARR framework and explains how each of its phases is grounded in that literature. Section 5 draws out the implications for platform design. Section 6 summarizes the argument and identifies directions for future empirical work.

2. Vocabulary Learning in Digital Language Platforms

2.1. Typical Vocabulary Learning Approaches in Language Apps

Digital language learning platforms have expanded considerably over the past decade in both reach and technical capability. Many now offer personalized feedback, adaptive content selection, and AI-driven conversational practice (Godwin-Jones, 2014; Lan, 2020). What has changed far less, however, is the underlying logic of vocabulary instruction. Across most major platforms, the approach follows a consistent pattern: words are introduced, matched to translations or images, and rehearsed through repetition until the learner achieves a target score.

This pattern is not difficult to explain. As Reinders and Wattana (2015) observed, digital language education has tended to prioritize activities that are easy to build, track, and scale over those that are pedagogically richer. Flashcard systems and multiple-choice quizzes fit that description well. They generate measurable outcomes and lend themselves to gamification. The concern, articulated at length by Nation (2022), is that what they generate may be a thin form of word knowledge: the ability to recognize a form and retrieve a rough equivalent, rather than the deeper understanding needed to use a word productively in real contexts.

Nation's (2022) distinction between shallow and deep vocabulary knowledge is worth dwelling on. Shallow knowledge amounts to recognizing a word when it appears

and matching it to a translation. Deep knowledge involves understanding how the word behaves across contexts, what it collocates with, and how to deploy it accurately in speech and writing. The success metrics most platforms rely on, including completion rates, daily streaks, and quiz scores, are designed around shallow knowledge. They measure recognition, not use.

2.2. Decontextualized Vocabulary Learning

One structural feature common to most digital vocabulary instruction is the presentation of words in isolation. Nation (2022) argued that genuine acquisition requires repeated exposure across varied contexts because it is through varied encounters that a learner builds up a full picture of how a word behaves. A learner who has reviewed a word dozens of times on a flashcard may still struggle to use it in conversation, not from a failure of effort, but because the flashcard encounters never supplied the contextual information that productive use depends on.

Webb (2008) showed that the depth of vocabulary knowledge grows through encounters that vary in both context and cognitive demand. Seeing the same word in the same format repeatedly may do relatively little to deepen understanding beyond the first encounter. Schmitt (2008) arrived at a similar conclusion, finding that how encounters are distributed over time matters as much as how many there are. Massed practice tends to produce weaker long-term retention than spaced exposure embedded in contexts that require learners to process a word's meaning rather than simply recognize its form.

2.3. Limitations of Screen-Based Interaction

The limitations described above are not only a matter of instructional method. They may also reflect something about the medium itself. Dede (2009) was among the first to make this argument clearly, drawing a contrast between the learner's relationship to a screen and to an immersive environment. On a screen, the learner is an observer of content. In an immersive environment, the learner is an inhabitant of a context. The difference is not cosmetic. It changes what kinds of experience are possible and, as a consequence, what kinds of learning can occur.

Mayer and Moreno (2003) established that the spatial arrangement of information affects how efficiently it is processed and retained. Lan (2020) and Godwin-Jones (2014) both noted that while language learning technology has diversified considerably, the screen has remained the default interface. Makransky and Lilleholt (2018) offered empirical support for the potential of alternatives: learners in immersive virtual reality environments showed stronger motivational engagement than those using screen-based tools, and motivation is among the more reliable predictors of long-term vocabulary growth. Taken together, these findings suggest that the screen is not a neutral surface for delivering instruction. It carries affordances and constraints that have shaped what digital vocabulary learning has been able to be.

This structural contrast can be more easily understood through the schematic representation shown in Figure 1. As illustrated in Figure 1, traditional approaches are structured around input and recognition, whereas the approach proposed in this paper organizes vocabulary learning as a cyclical process involving interaction, retrieval, and reuse.

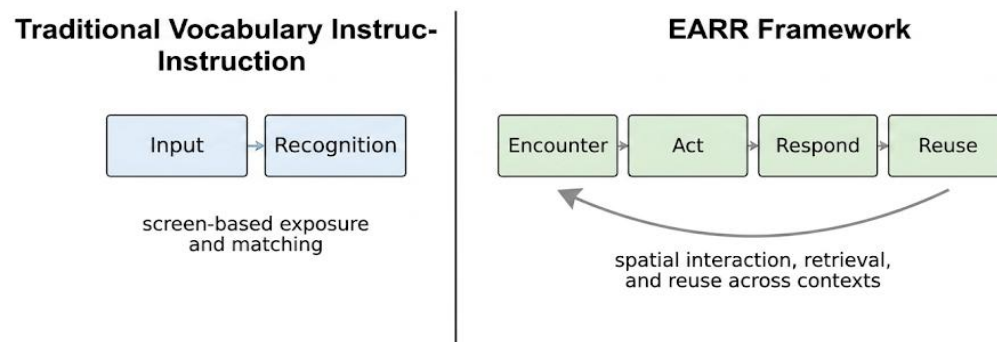


Figure 1. Comparison between traditional vocabulary instruction and the EARR framework.

3. Theoretical Foundations of Experiential Vocabulary Learning

3.1. Vocabulary Acquisition and Contextual Learning

The research on vocabulary acquisition converges on a view that runs counter to how most digital platforms operate. Nation (2022) documented extensively that word knowledge is not fixed at a single encounter but develops gradually through repeated exposure across varied contexts. Webb (2008) traced this development along a continuum, from shallow receptive recognition to deeper productive command, and showed that progression along that continuum depends on the quality of encounters as much as their quantity. Schmitt (2008) made the same point from the perspective of instructed learning, arguing that spaced, contextually embedded exposures are more effective than concentrated drills.

Roediger and Karpicke (2006) added a further dimension to this picture by demonstrating that actively recalling information produces substantially stronger long-term retention than passive review. Cepeda et al. (2006) showed that distributing retrieval practice over time amplifies this effect considerably. The practical implication is that what matters most in vocabulary instruction may not be how much time a learner spends with a word but the nature of that time: whether it involves active engagement, distributed retrieval, and contextual variation.

This may raise the question of how such conditions might be created within digital learning environments. Recent studies have suggested that immersive virtual reality environments can support vocabulary acquisition by enabling learners to engage in language in context-rich, interactive settings (Alfakil, 2024).

3.2. Embodied Cognition and Action-Based Language Learning

Mainstream accounts of vocabulary instruction have tended to treat language as a cognitive and symbolic matter, something that happens in the mind and is expressed through the mouth or the hand. Embodied cognition takes a different view. It holds that meaning is not stored as abstract propositions but is grounded in sensorimotor experience, in the perceptual and motor states that accompany encounters with the world. Understanding a word, on this account, is not simply a matter of retrieving a definition. It involves a partial simulation of the experiences associated with what the word refers to (Barsalou, 2008).

Barsalou (2008) developed this argument in detail, proposing that conceptual knowledge consists of multimodal simulations constructed from the sensory, motor, and affective traces of prior experience. If this account is correct, then vocabulary instruction that bypasses the body may be working against the mechanisms through which meaning is formed. Wilson (2002) extended the argument by showing that even the kind of cognition that seems most detached from the body, such as reading or mental imagery, draws on sensorimotor systems that were originally engaged in direct interaction with the world.

Glenberg and Kaschak (2002) provided a well-known empirical demonstration of this through the action-sentence compatibility effect, showing that understanding sentences about action activates the motor systems involved in performing those actions. More recently, Lindgren and Johnson-Glenberg (2013) found that learners who enacted scientific concepts through physical movement in mixed-reality environments retained more than those who observed the same content without moving. For vocabulary instruction, the implication is that words may be most durably learned when the learning context engages the body as well as the mind.

3.3. Situated Learning and Spatial Context

Situated learning theory, as Lave and Wenger (1991) developed it, holds that knowledge is not a portable abstraction that a learner carries from one context to another but something that is tied to the situations in which it is acquired and used. Learning, in this view, is a process of becoming progressively capable of participating in the practices of a particular community or setting. It is driven not by the transmission of information but by engagement with authentic tasks that have genuine purpose in a real context.

For vocabulary learning, this perspective has a direct implication. A word encountered in a meaningful spatial context is not merely easier to remember. It is learned in a different way. When a learner comes across the word “shelf” while reaching for an object on one, the word arrives already connected to a spatial relationship, a physical action, and a situation in which its meaning matters. Brown et al. (1989) made a related point in arguing that knowledge is fundamentally situated and that effective learning environments should reflect the conditions under which knowledge will eventually be used. Most current language platforms do not approach this standard. Words appear on screens with no surrounding environment, no objects to interact with, and no purpose that gives meaning beyond the next quiz item.

Recent reviews have further suggested that immersive, interactive environments may support the kind of contextualized language use emphasized in situated learning theory (Huang et al., 2021). These environments may allow learners to engage in settings that more closely resemble real-world contexts, which may help bridge the gap between exposure and use.

3.4. Experiential Learning Theory

Dewey (1986) argued that experience is not a backdrop to learning but its primary medium. What distinguishes educative experience from mere activity, in his account, is whether the learner reflects on what has happened, draws meaning from it, and carries that meaning into subsequent situations. Kolb (2014) gave this insight a more formal structure through his experiential learning cycle, which describes learning as moving through four stages: concrete experience, reflective observation, abstract conceptualization, and active experimentation. Genuine understanding, on Kolb’s account, requires the full cycle.

Applied to vocabulary acquisition, this framework points to something most digital platforms do not provide. A learner who sees a word, perhaps hears it pronounced, and answers a multiple-choice question has passed through something like the first stage. What follows in a typical app is more of the same: another word, another prompt, another answer. There is no action that puts the word to use, no feedback that connects it to a meaningful outcome, and no later context in which it must be retrieved under different conditions. Reinders and Wattana (2015) found that game-based language tasks incorporating these experiential elements produced stronger vocabulary outcomes than conventional drill activities, lending empirical weight to what the theory suggests. Durable learning, it appears, depends on purposeful action, meaningful feedback, and varied re-engagement over time.

4. The EARR Framework for Spatial Experiential Vocabulary Learning

4.1. Spatial Learning Environments in Digital Platforms

Spatial learning environments are settings in which the physical or virtual arrangement of space contributes meaningfully to the learning experience. Mayer and Moreno (2003) established that how information is spatially organized affects how it is processed and retained. Uttal et al. (2013) showed through meta-analysis that spatial reasoning is trainable and transfers across academic domains, raising the possibility that navigating and interacting with spatially arranged objects may recruit cognitive processes that support deeper encoding.

The potential of immersive environments for language learning has attracted growing attention. Makransky and Lilleholt (2018) found that learners in virtual reality environments showed stronger motivational engagement than those in conventional screen-based settings. Lan (2020) and Godwin-Jones (2014) both documented this turn toward immersive and interactive contexts in the research literature. At the same time, Parong and Mayer (2018) found that immersive environments did not consistently outperform traditional instruction and, in some conditions, produced weaker outcomes. This is an important qualification. Spatial environments are not inherently superior to conventional ones. Their value depends on how they are designed and whether spatial features are genuinely connected to learning objectives rather than added for their own sake.

4.2. Object-Based Vocabulary Learning

A consistent finding across vocabulary research is that words connected to concrete, manipulable objects tend to be acquired more readily than abstract terms. Barsalou (2008) offered a theoretical explanation grounded in embodied cognition: object-related words are well suited to experiential learning because the objects themselves invite perceptual and motor engagement, producing richer conceptual representations than those formed through symbolic study alone. A word learned through interaction with its referent is theoretically expected to be encoded with more associative depth than one learned from a definition, though direct evidence for this in digital vocabulary learning contexts remains limited (Parong & Mayer, 2018).

Glenberg and Kaschak (2002) extended this reasoning by showing that understanding words about physical objects involves simulating the actions those objects afford. Plass et al. (2015) found empirical support in the context of multimedia learning, showing that environments incorporating interactive object manipulation supported vocabulary retention more effectively than static image and word pairings. The interactive quality of the encounter, not merely its visual richness, appears to be what matters. For platform designers, this suggests that placing learners near objects is less important than giving them something meaningful to do with them.

4.3. Action-Based Interaction and Vocabulary Retrieval

Contextual encounter is a necessary starting point for vocabulary acquisition, but it is not sufficient on its own. What converts an encounter into durable learning depends substantially on what the learner does in response. When a learner interacts with an object in a spatial environment, performing an action that the target vocabulary describes or refers to, the word is encoded not only as a linguistic form but also as part of a sensorimotor experience. That kind of encoding tends to support stronger retrieval than one built on visual exposure alone, consistent with embodied cognition accounts, though the extent to which this holds in digital spatial environments specifically remains an open empirical question (Glenberg & Kaschak, 2002).

Wilson (2002) suggested that memory for language is not stored independently of the bodily states that accompanied its acquisition. Retrieving a word may involve reconstructing something of the original experience, drawing on motor and perceptual traces that remain available even when the learner is no longer in the learning environment. Lindgren and Johnson-Glenberg (2013) found evidence for this in the context of augmented reality, where learners who had enacted concepts physically retained them more strongly than those who had only observed. If retrieval works in part through embodied memory, then instruction that creates richer embodied encoding during initial learning may yield stronger retrieval later. This points toward an instructional principle: the actions learners perform in a spatial environment should not be incidental to the vocabulary being taught. They are most likely to be effective when they make the meaning of target words physically and functionally concrete.

4.4. The EARR Framework

The theoretical perspectives reviewed in this paper converge on a view of vocabulary learning that differs in important ways from what current digital platforms typically provide. Building on the theoretical perspectives discussed above, this paper proposes the EARR framework as a structured four-phase model for vocabulary acquisition in spatially situated digital environments. The framework is intended as a theoretically grounded proposal rather than an empirically validated model. Its claims about learning outcomes await direct testing under controlled conditions. As illustrated in Figure 2, the EARR framework consists of four phases: Encounter, Act, Respond, and Reuse, which together form a cyclical learning process.

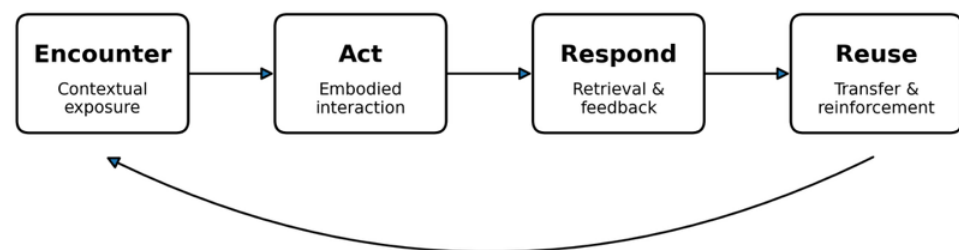


Figure 2. The EARR Framework for Spatial Experiential Vocabulary Learning.

It is worth situating EARR in relation to prior frameworks. Kolb's (2014) experiential learning cycle is the most obvious reference point, and the two frameworks share an emphasis on active engagement and iterative experience. The differences are substantial, however. Kolb's cycle is a general model of experiential learning that applies across domains and does not address the specific mechanisms of vocabulary acquisition. EARR, by contrast, is developed specifically for vocabulary learning in spatially situated digital environments, and each of its phases is grounded in research on how words are encoded, retained, and retrieved. Existing work on immersive language learning has focused primarily on motivational and affective outcomes (Makransky & Lilleholt, 2018). EARR is oriented toward the cognitive and sensorimotor mechanisms through which spatial, object-based interaction may support durable word knowledge, drawing on embodied cognition and situated learning theory in ways that prior frameworks have not brought together for this purpose.

The first phase, Encounter, provides the perceptual and situational grounding for what follows. In a spatially situated environment, a learner comes across a word because they are in a place where its referent exists and where its meaning is available from the surrounding context. Following Lave and Wenger (1991), this situational arrival is quali-

tatively different from seeing a word introduced on a screen. The word comes with a setting already attached to it.

Act, the second phase, is where embodied learning takes place. The learner does something with the word, like picking up the object it refers to, performing the action it describes, or using it to accomplish a goal within the environment. Drawing on Glenberg and Kaschak (2002) and Lindgren and Johnson-Glenberg (2013), this kind of action may recruit the same motor systems involved in the described activity, encoding the word with sensorimotor depth that passive reading or listening cannot produce.

Response is the retrieval and feedback phase. The environment reacts to what the learner has done, providing a signal that connects the word to a meaningful consequence. This operationalizes the retrieval practice principles documented by Roediger and Karpicke (2006) in a spatially grounded context. The learner is not simply exposed to the word again. They receive feedback that makes its use consequential in a way that a correct answer on a quiz cannot fully replicate.

The fourth phase, Reuse, asks the learner to encounter and use the word again in a different context. Following Kolb (2014), learning may not be complete until understanding has been applied in a new situation requiring genuine retrieval rather than recognition. The spatial environment is what holds the four phases together and gives them coherence. A learner who moves through all four phases repeatedly, in varied settings, may reach a point where a word is no longer retrieved with effort but is simply available as a tool. That represents a meaningfully higher standard than what recognition-based instruction is typically designed to achieve. These four phases constitute a structured framework for both designing and evaluating vocabulary learning in spatial environments.

4.5. Illustrative Example of the EARR Framework

In order to make the EARR framework more concrete, a brief example of how a single vocabulary item may be learned within a spatially situated environment is presented in this section.

When a learner enters a virtual living room and encounters a lamp placed on a table, the word “lamp” is not introduced as an isolated item but appears in direct relation to the object within its environment. It is supported by auditory input and minimal text cues. In this initial encounter, the learner does not only see the word but also encounters it as part of a meaningful situation.

The learner is then guided to interact with the object. They may turn on the lamp, move closer to it, or adjust its position. Through these actions, the learner engages directly with what the word refers to, allowing the word to be linked to a physical experience rather than remaining a detached form. This stage reflects the role of action in strengthening the connection between language and experience.

Following this interaction, the environment provides feedback that requires the learner to retrieve and use the word. The learner may be prompted to describe the action they have just performed. For example, the learner produces a simple sentence such as “I turned on the lamp”. At this point, the word must be actively brought to mind and applied in context, moving the learner beyond recognition toward meaningful use.

At a later stage, the word reappears in a different setting. The learner may encounter a similar object in another room or be required to use the word in a new task involving lighting. Because the situation has changed, the learner must draw on prior experience to use the word again under different conditions, strengthening the flexibility and durability of their knowledge.

This example shows how vocabulary learning can unfold as a sequence of related experiences in which meaning is built through interaction, reinforced through use, and

extended across contexts. Through repeated engagement of this kind, the word gradually becomes available for use rather than remaining something that is only recognized.

The framework applies equally across different spatial contexts and vocabulary types. A second example scenario illustrates this in a virtual grocery store. Upon entering the space, the learner encounters the word “aisle” as they navigate between shelves, with the word appearing alongside the object it refers to and supported by a brief audio prompt. In the Act phase, the learner is asked to find a specific item by following spoken directions that use the word in context. For example, “Please find the cereal in the third aisle on your left”. The environment responds when the learner reaches the correct location, confirming the action and reinforcing the word’s spatial meaning. In a later task, the same word reappears in a different form. The learner must give directions to another character, requiring them to retrieve and produce the word independently rather than simply respond to it.

A third example scenario extends the framework to a professional context. In a virtual office environment, the learner encounters vocabulary associated with workplace tasks. The word “agenda” first appears during a meeting room scene, where the learner sees a printed agenda on the table and hears it referenced in a spoken prompt. In the Act phase, the learner is asked to distribute copies of the agenda to colleagues seated around the table, performing a purposeful action that connects the word to its functional role. Feedback follows in the form of a colleague’s response acknowledging the action. The Reuse phase then asks the learner to prepare a simple agenda for a new meeting, requiring them to apply the word in a productive rather than a receptive context.

These example scenarios illustrate that the EARR framework is not tied to any single environment or register. What remains consistent across contexts is the underlying structure where the word arrives with a situation, the learner acts on it, the environment responds, and the word reappears under different conditions. The spatial setting changes, but the learning logic does not.

5. Implications for Language Learning Platforms

The design principles proposed in this section follow directly from the theoretical perspectives developed in Section 3, and the connection between them is worth making explicit. Situated learning theory (Lave & Wenger, 1991) implies that learning environments should present vocabulary within authentic, purposeful contexts rather than in isolation, which translates directly into the first design principle that the environment should do pedagogical work. Embodied cognition (Barsalou, 2008; Glenberg & Kaschak, 2002) implies that meaning is constructed through sensorimotor engagement, which translates into the requirement that learners act on objects rather than observe them. Experiential learning theory (Kolb, 2014) implies that learning is completed only when the full cycle of encounter, action, feedback, and reuse has been traversed, which translates into the principle that all four phases of the EARR cycle must be consistently delivered. The cognitive load considerations raised by Mayer and Moreno (2003) and Parong and Mayer (2018) serve as a constraint on all three principles. Spatial richness, object interaction, and cycle delivery must each be calibrated so that the cognitive demands of the environment do not exceed the learner’s available processing capacity. A spatially rich environment that overwhelms working memory is not an implementation of these principles but a departure from them.

5.1. From Screen-Based Learning to Spatial Learning

The argument developed in this paper has a specific implication for how language learning platforms are conceived. The dominant model positions the learner as an observer who receives and responds to content delivered on a screen. Dede (2009) framed

the alternative in terms of a shift in the learner's fundamental relationship to the learning environment. In a spatial setting, the learner is an inhabitant of a context rather than a viewer of a display. This is not a distinction of degree. An inhabitant encounters situations, makes decisions, performs actions, and experiences their consequences. Those are closer to the conditions under which language is actually used, and, according to situated learning theory, the conditions under which it is most durably learned.

Godwin-Jones (2014) and Lan (2020) both noted growing recognition in the field that language learning technology may need to move toward environments that support genuine, purposeful interaction rather than simulated engagement. Uttal et al. (2013) added a cognitive perspective, showing that engagement with spatially rich environments may develop reasoning capacities that transfer across academic domains. The shift from screen-based to spatial learning is, in this sense, not primarily a technological matter. It is a pedagogical one, requiring a different set of assumptions about what learning environments are for and how they can be designed to support the conditions vocabulary acquisition appears to require.

5.2. Advantages of Object-Based Experiential Vocabulary Learning

The potential advantages of spatial experiential vocabulary learning are cognitive rather than motivational, though the two are not entirely separate. The most fundamental concern is the quality of encoding. Nation (2022) and Webb (2008) both showed that what drives progress along the vocabulary knowledge continuum is not the number of exposures but the richness and variety of the encounters. Object-based interaction in a spatial environment may provide exactly that: an encounter encoded with sensory, motor, and situational information that a flashcard or a multiple-choice prompt cannot supply. Plass et al. (2015) found support for this, showing that interactive object manipulation in multimedia environments produced stronger retention than static image and word pairings.

A second potential advantage concerns the accessibility of words during retrieval. Words learned through embodied interaction may be connected to a broader network of associations that provide more retrieval pathways than words learned through visual pairing alone. The object is not a cue for the word; it is the context in which the word first acquired its meaning. That said, Parong and Mayer (2018) are a useful reminder that spatial richness does not translate automatically into learning benefit. Environments that present too much at once can impede learning rather than support it. Design quality is likely to be at least as important as the choice of medium.

5.3. Implications for Future Language Learning Platforms

Three design principles follow from the argument developed in this paper. The first is that the learning environment should do genuine pedagogical work. Many attempts to build richer digital environments have added spatial or interactive features without reconsidering the instructional logic that underlies them (Godwin-Jones, 2014; Lan, 2020). A spatially designed vocabulary learning environment should be one in which objects, locations, and tasks are chosen because they make target words meaningful and memorable, not because they are visually engaging or technically impressive.

The second principle concerns cognitive load. Spatial richness and cognitive demand are in tension, and that tension requires careful management. Parong and Mayer (2018) showed that immersive environments can overwhelm the cognitive system when they present too much at once. Makransky et al. (2019) found that high-immersion conditions can paradoxically reduce learning by increasing extraneous load. Good design in this context means calibrating the complexity of the spatial environment to the learning demands of the vocabulary being taught so that the spatial features direct attention toward meaning rather than away from it. The third principle concerns the integrity of the EARR cycle. A

platform built around this framework should ensure that learners move through all four phases, not just the first two. Encounter and Act may be relatively straightforward to design for. Respond and Reuse require more deliberate attention: feedback must be timely and consequential, and vocabulary must reappear across varied contexts and tasks over time. A platform that delivers the first two phases without the latter two may produce richer initial encoding than a flashcard system while still falling short of the spaced, varied retrieval practice that supports durable acquisition.

Adopting the EARR framework in practice presents challenges that are worth acknowledging directly. Building spatially rich learning environments requires considerably more design investment than producing flashcard-based content. Educators working without dedicated development support may find the resource demand prohibitive, particularly in contexts where content must be regularly updated or adapted for different learner groups. Platforms that offer integrated authoring tools can reduce some of this burden, but they introduce their own learning curve and do not resolve the underlying design challenges of ensuring that spatial features serve learning objectives rather than simply adding visual complexity.

A further challenge concerns proficiency calibration. The EARR framework assumes that learners can engage meaningfully with spatial tasks, but the appropriate level of linguistic and cognitive demand will vary considerably across proficiency levels, games, and learning contexts. An environment designed for adult beginners in a workplace setting may look quite different from one designed for younger learners acquiring everyday vocabulary. Designers will need to make deliberate decisions on how to scaffold the Encounter and Act phases for learners who may lack the linguistic resources to engage with complex spatial prompts, without reducing those phases to the kind of simplified recognition tasks the framework is intended to move beyond.

Finally, ensuring that all four phases of the EARR cycle are consistently delivered requires sustained attention across the full arc of a learning sequence. Encounter and Act are relatively tractable design problems. Respond requires feedback mechanisms that connect word use to meaningful consequences rather than a binary correct or incorrect signal. Reuse requires a content architecture that tracks what vocabulary a learner has encountered and surfaces it again across different contexts and tasks over time. These are not insurmountable problems, but they are non-trivial, and platforms that address them only partially may capture some of the benefits of spatial experiential learning while falling short of what the full cycle is designed to produce.

The OnTwins platform (www.ontwins.com) offers one example of how these principles might be realized in practice. OnTwins is an integrated content generation and deployment platform that combines document-based knowledge, spatial environment definition, and AI-driven scripting to produce immersive educational content across multiple execution environments, including VR headsets, AR systems, mobile devices, and desktop interfaces. The overall architecture of the platform is illustrated in Figure 3.

The platform operates as a coordinated pipeline. Educational materials and natural language prompts enter through a user input interface, where learning objectives, difficulty levels, and interaction styles are specified. A spatial definition module then establishes the virtual environment, drawing on 3D assets and interface scripts that determine how learners engage with objects in the space. The logical structure of each learning sequence is governed by a script definition module, which manages learning stages, conditional branching, and feedback delivery. These components come together in a central authoring tool that coordinates AI-generated content with human review and refinement. The resulting content is stored in a repository supporting version control and reuse and rendered through Unity, Unreal, or WebGL depending on the target device and context.

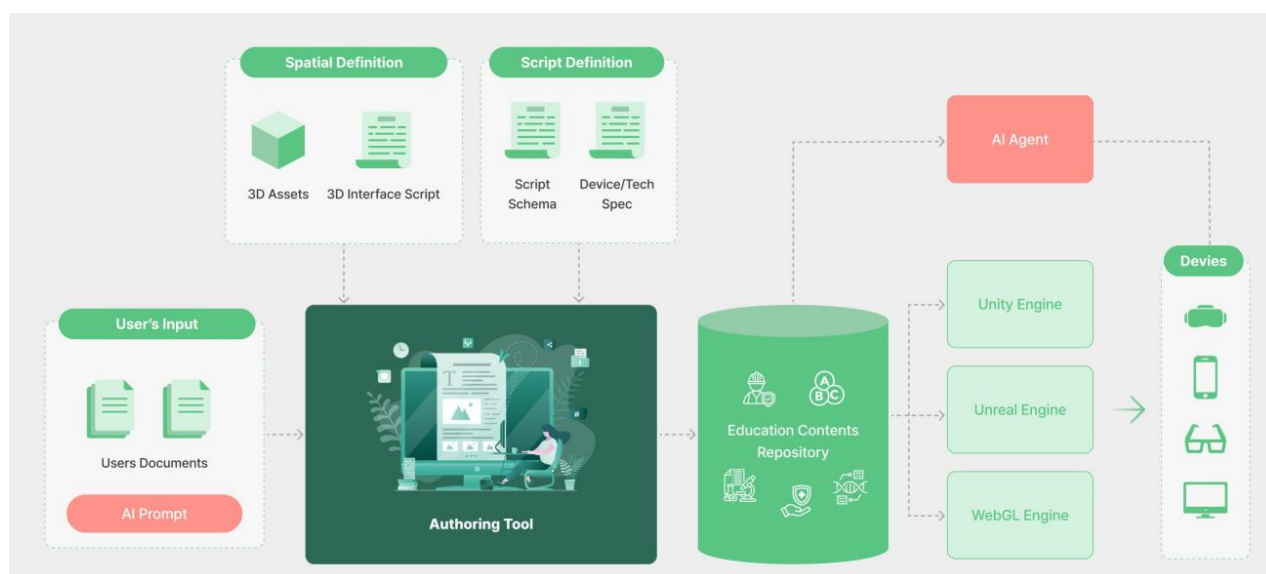


Figure 3. The OnTwins platform pipeline.

Figure 4 illustrates how this sequence is realized within the OnTwins environment, using the target word “lamp” as an example. In panel (a), the learner encounters the word in a virtual living room where the lamp is present as a physical object, with the word displayed and read aloud simultaneously, grounding it in its spatial context before any task is introduced. The learner is then given a purpose: panel (b) shows the prompt “It’s a bit dark. Turn on the lamp”, which asks the learner to do something with the word rather than simply recognize it. Panel (c) shows the environment responding to the learner’s action with immediate confirmation, making the consequence of using the word visible within the space. In panel (d), the learner is asked “Is the lamp on?”, turning what began as an instruction into a question that requires the learner to bring the word back to mind on their own terms.

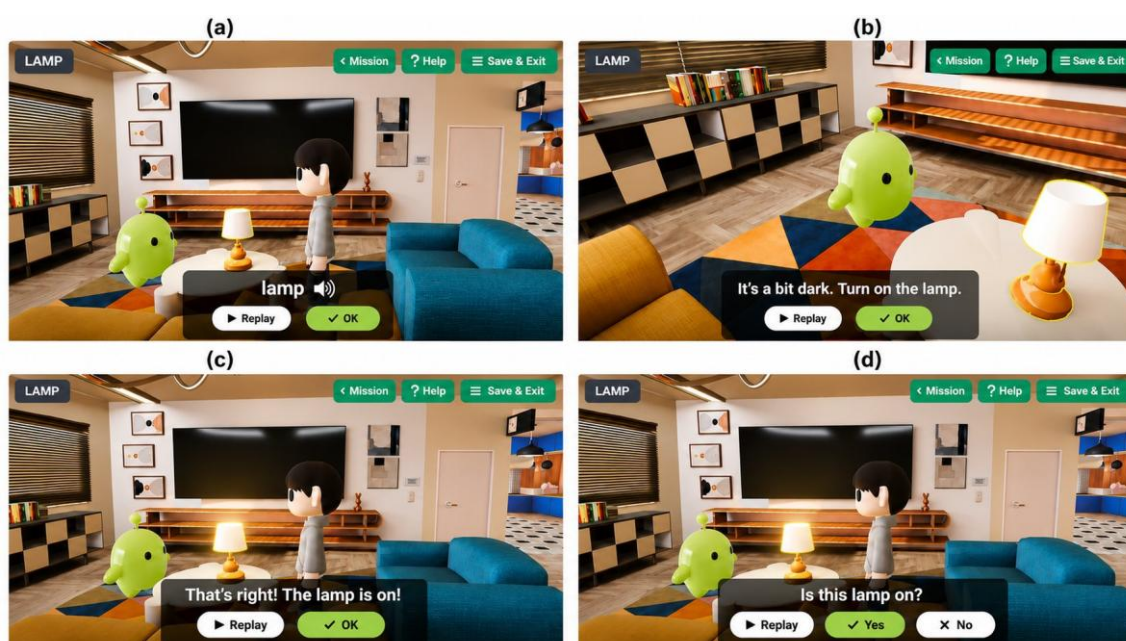


Figure 4. Screenshots from the OnTwins illustrating the EARR (target word “lamp”): (a) Encounter: The learner approaches the lamp and the word “lamp” is presented in text and sound. (b) Act: The helper encourages the learner to turn on the lamp. (c) Respond: Immediate feedback is provided

after the learner turns on the lamp. (d) Reuse: The word “lamp” is reused in a new prompt to assess understanding in context.

For the purposes of this paper, what matters is how this architecture relates to the conditions the EARR framework identifies as important. The spatial definition module and 3D object library support the Encounter phase by embedding vocabulary within object-rich environments rather than introducing words on a flat screen. The interaction logic in the interface scripts supports Act, giving learners something purposeful to do with the objects and situations that target words refer to. The feedback structures in the script schema address Respond, connecting word use to a meaningful consequence rather than a simple correct or incorrect signal. And the content repository, through its support for scenario variation and reuse, provides the infrastructure for the Reuse phase, ensuring that vocabulary reappears across different contexts over time. OnTwins is not a test of the EARR framework. It is an attempt to build a platform whose architecture takes seriously the learning conditions the framework describes.

The script definition module plays a particularly important role in operationalizing the Respond and Reuse phases of the EARR cycle. Within the Respond phase, the module governs the feedback logic that connects a learner’s action to a meaningful consequence. Rather than delivering a binary correct or incorrect signal, the script schema allows designers to specify conditional response pathways. If a learner selects the correct object, the environment responds with a contextually appropriate confirmation, and if the learner’s response is incorrect or incomplete, the schema triggers a scaffolded prompt that redirects attention without simply repeating the original instruction. The feedback architecture is what distinguishes Respond from mere assessment. In the Reuse phase, the script definition module coordinates with the content repository to determine when and how a previously encountered word reappears. The module tracks the vocabulary a learner has encountered across a session and surfaces the same word in a new scenario with different task demands, ensuring that reuse requires genuine retrieval rather than simple recognition of a familiar prompt.

Dede (2009) suggested that the most significant obstacle to realizing the potential of immersive learning environments is not technological but conceptual. The platforms most likely to advance vocabulary instruction are those built by designers who understand what acquisition requires and are willing to let that understanding guide their decisions, even when it runs against simpler alternatives.

6. Conclusions

This paper has argued that the dominant approach to vocabulary instruction in digital language learning platforms is unlikely to produce the kind of durable word knowledge that acquisition research describes. Flashcards, translation drills, and recognition-based quizzes are well-optimized for measurability and engagement, but they tend to present words in isolation, demand recognition rather than use, and provide little of the contextual variety and active retrieval that the literature consistently identifies as important for lasting learning.

In response, the paper has proposed spatial experiential vocabulary learning as an alternative and introduced the EARR framework as a structured approach to its design. The framework draws on four bodies of research: vocabulary acquisition research, embodied cognition, situated learning theory, and experiential learning theory. Each provides a different piece of the argument. Vocabulary research establishes what durable acquisition requires. Embodied cognition explains why physical engagement with language may matter. Situated learning theory shows why context is not incidental but constitutive of what is learned. Experiential learning theory provides a model of the cycle through

which knowledge is built and consolidated. EARR integrates these perspectives into a framework specific to vocabulary acquisition in spatially situated digital environments, distinguishing it from prior models that have addressed these themes separately or in different domains.

The most immediate priority for future research is empirical. The theoretical case for spatial experiential vocabulary learning is supported by converging evidence from multiple traditions, but the direct effects of EARR-based instruction on vocabulary outcomes have not yet been tested under controlled conditions. The framework makes specific predictions that are, in principle, testable. It predicts that learners who move through all four phases of the EARR cycle will show stronger long-term retention than those exposed to equivalent vocabulary through conventional screen-based methods; that the Act phase will contribute to encoding depth in ways that passive encounter alone cannot; and that the Reuse phase will support retention durability beyond what massed practice produces. Each of these predictions corresponds to a body of existing evidence, but none has been tested directly in the context of spatially situated digital vocabulary instruction. Several research designs would be appropriate for this work. Pre-post vocabulary assessments comparing EARR-based instruction with matched conventional instruction across equivalent learner populations and vocabulary sets would provide an initial test of the framework's central claims. Delayed post-tests, administered weeks or months after instruction, would be particularly informative given the framework's emphasis on durable acquisition rather than short-term retention gains. Eye-tracking studies could examine how spatial environments direct learner attention during the Encounter phase, testing whether object-grounded encounters produce different patterns of attention than screen-based word presentation. Think-aloud protocols and behavioral logs could shed light on how learners engage with the Act and Respond phases, providing process-level evidence to complement outcome measures. Equally important is research on which design features of spatial environments support learning and which impede it, building on the work of Parong and Mayer (2018) and Makransky et al. (2019). Questions about how the EARR framework applies across different proficiency levels, learning contexts, and target vocabulary types also remain open. The framework offered here is intended as a starting point for that work, not a conclusion.

One limitation of the EARR framework as currently formulated deserves explicit acknowledgement. The framework is most naturally suited to vocabulary that refers to concrete, spatially situated objects and actions. Words like “lamp”, “aisle”, or “agenda” have clear physical references that can be represented in a spatial environment and interacted with directly. Abstract vocabulary, including terms such as “justice”, “ambiguity”, or “perseverance”, does not lend itself to the same kind of object-based encounter and action. This does not mean the framework is inapplicable to abstract vocabulary, but it does mean that the Encounter and Act phases would need to be designed differently, perhaps through metaphorical spatial representation, narrative context, or dialog-based interaction rather than direct object manipulation. Future work should address how the EARR cycle can be adapted for abstract vocabulary types and whether the embodied and spatial grounding that supports acquisition of concrete words translates, in modified form, to more abstract registers.

Author Contributions: Conceptualization, H.K., S.C. and J.W.; Methodology, H.K., S.C. and J.W.; Software, S.C. and J.W.; Resources, J.W.; Data curation, J.W.; Writing—original draft, H.K.; Writing—review & editing, S.C. and J.W.; Supervision, H.K.; Project administration, S.C.; Funding acquisition, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AR	Augmented Reality
EARR	Encounter, Act, Respond, and Reuse
VR	Virtual Reality

References

- Alfadil, M. (2024). Immersive virtual reality: A novel approach to second language vocabulary acquisition in K-12 education. *Sensors*, 24(22), 7185.
- Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, 59, 617–645. <https://doi.org/10.1146/annurev.psych.59.103006.093639>.
- Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18(1), 32–42. <https://doi.org/10.3102/0013189x018001032>.
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3), 354–380. <https://doi.org/10.1037/0033-2909.132.3.354>.
- Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, 323(5910), 66–69. <https://doi.org/10.1126/science.1167311>.
- Dewey, J. (1986, September). Experience and education. In *The educational forum* (Vol. 50, pp. 241–252). Taylor & Francis Group.
- Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, 9(3), 558–565. <https://doi.org/10.3758/BF03196313>.
- Godwin-Jones, R. (2014). Games in language learning: Opportunities and challenges. *Language Learning & Technology*, 18(2), 9–19.
- Huang, X., Zou, D., Cheng, G., & Xie, H. (2021). A systematic review of AR and VR enhanced language learning. *Sustainability*, 13(9), 4639.
- Kolb, D. A. (2014). *Experiential learning: Experience as the source of learning and development*. FT Press.
- Lan, Y. J. (2020). Immersion into virtual reality for language learning. In *Psychology of learning and motivation* (Vol. 72, pp. 1–26). Academic Press.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge University Press.
- Lindgren, R., & Johnson-Glenberg, M. (2013). Emboldened by embodiment: Six precepts for research on embodied learning and mixed reality. *Educational Researcher*, 42(8), 445–452. <https://doi.org/10.3102/0013189x13511661>.
- Makransky, G., & Lilleholt, L. (2018). A structural equation modeling investigation of the emotional value of immersive virtual reality in education. *Educational Technology Research and Development*, 66(5), 1141–1164. <https://doi.org/10.1007/s11423-018-9581-2>.
- Makransky, G., Terkildsen, T. S., & Mayer, R. E. (2019). Adding immersive virtual reality to a science lab simulation causes more presence but less learning. *Learning and Instruction*, 60, 225–236. <https://doi.org/10.1016/j.learninstruc.2017.12.007>.
- 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.
- Nation, I. S. P. (2022). *Learning vocabulary in another language*. Cambridge University Press.
- Parong, J., & Mayer, R. E. (2018). Learning science in immersive virtual reality. *Journal of Educational Psychology*, 110(6), 785–797. <https://doi.org/10.1037/edu0000241>.
- Plass, J. L., Homer, B. D., & Kinzer, C. K. (2015). Foundations of game-based learning. *Educational Psychologist*, 50(4), 258–283. <https://doi.org/10.1080/00461520.2015.1122533>.
- Reinders, H., & Wattana, S. (2015). Affect and willingness to communicate in digital game-based learning. *ReCALL*, 27(1), 38–57. <https://doi.org/10.1017/S0958344014000226>.

- Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>.
- Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research*, 12(3), 329–363. <https://doi.org/10.1177/1362168808089921>.
- Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, 139(2), 352–402. <https://doi.org/10.1037/a0028446>.
- Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*, 30(1), 79–95. <https://doi.org/10.1017/S0272263108080042>.
- Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4), 625–636. <https://doi.org/10.3758/BF03196322>.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.