TEFL Journal

Vo 4, No 2 , 2026

The Effect of Multimodal Instruction on Vocabulary Learning and Retention in Iraqi EFL Classrooms

Rusul Esmaeel Sakran

Article Info

Article History

Received:
1 March 2026

Revised
1 May 2026

Accepted:
1 July 2026

Published
27 July 2026

Keywords

Multimodal instruction, vocabulary learning, vocabulary retention, Iraqi EFL learners, multimedia learning

Abstract

Vocabulary knowledge remains a fundamental component of second language competence, yet Iraqi EFL learners face considerable challenges in acquiring and retaining lexical items due to limited exposure to authentic English and predominant reliance on traditional text-based instruction. This quasi-experimental study examined the impact of multimodal instruction on vocabulary learning and retention among 60 intermediate Iraqi EFL learners randomly assigned to an experimental group (n = 30) receiving instruction integrating visual, auditory, and textual materials, or a control group (n = 30) taught through conventional text-based methods. While extensive international research has documented the benefits of multimodal approaches for vocabulary acquisition, empirical evidence from the Iraqi EFL context, characterized by resource constraints, examination-driven pedagogy, and limited authentic language exposure, remains conspicuously absent. This study addresses this gap by drawing upon Dual Coding Theory (Paivio, 1971), Cognitive Theory of Multimedia Learning (Mayer, 2014), and Depth Processing Hypothesis (Craik & Lockhart, 1972) to investigate whether multimodal instruction yields superior learning outcomes in this under-researched context. Both groups completed a 30-item vocabulary test before intervention, immediately after, and two weeks later. One-way ANCOVA controlling for pre-test scores revealed that the multimodal group demonstrated significantly better performance on both the immediate post-test ($F(1,57) = 40.62$, $p < .001$, $\eta^2 = .413$) and the delayed post-test ($F(1,57) = 62.31$, $p < .001$, $\eta^2 = .521$), with large effect sizes indicating substantial practical significance. These findings empirically confirm that the simultaneous presentation of written definitions, visual illustrations, audio recordings, and video clips creates redundant memory traces and facilitates deeper semantic processing, thereby enhancing both initial vocabulary acquisition and long-term retention. The results provide strong theoretical validation for multimodal learning principles in a resource-limited EFL context and offer practical guidance for curriculum developers, material designers, and classroom practitioners seeking evidence-based strategies to improve vocabulary instruction in Iraqi educational settings.

1. Introduction

Vocabulary knowledge is a cornerstone of second-language competence, providing learners with the ability to understand input and to convey meaning in various skill areas (Nation, 2001; Schmitt, 2010). In the absence of a sufficiently large and highly processed vocabulary, learners fail to read authentic texts, follow spoken discourse, and produce coherent speech and writing (Wilkins, 1972). The significance of vocabulary is particularly relevant in English as a Foreign Language (EFL) contexts where exposure to the target language is often only possible in formal or structured formats and where the sheer quantity of words to be learned presents a continuing problem for both learners and teachers (Nation, 2006).

In Iraq, English language teaching has historically been teacher-centred, examination-oriented, and characterized by practices that limit authentic language use outside the classroom, with English often perceived merely as a test subject rather than a communicative tool (Saeed, 2015). The traditional approach, which emphasizes rote memorisation and translation, frequently causes students to struggle with vocabulary acquisition and retention (Al-Marzouqi, 2020). Public schools lack resources such as libraries, audio-visual tools, and additional materials to engage learners effectively, while classrooms suffer from large class sizes, noise pollution, and inadequate infrastructure (Saeed, 2015). These challenges are compounded by an unstable security situation that has resulted in additional interruptions to the educational process and decreased student motivation (Saeed, 2015). Such conditions render traditional text-centred vocabulary instruction largely ineffectual, as it fails to provide learners with the rich contextualised language input necessary for meaningful processing and long-term retention (Al-Marzouqi, 2020).

Given these contextual limitations, educational researchers have increasingly explored multimodal strategies as a means to overcome the shortcomings of traditional teaching by integrating visual, auditory, and textual information to achieve higher learning outcomes. In multimodal instruction, presenting content through various sensory modalities enhances cognitive processing and improves sustained retention (Paivio, 1971; Mayer, 2014). When vocabulary learning incorporates multimodal resources, such as images, video, audio recordings, and written annotations, learners gain access to a broader range of contextual elements and diverse retrieval pathways. This approach not only supports vocabulary acquisition but also promotes long-term retention (Chun & Plass, 1996; Yanguas, 2009). Additionally, it is believed to reduce cognitive load, boost motivation, and foster meaningful engagement with lexical items (Sweller, 1988; Mayer, 2014).

Despite the well-documented merits of multimodal instruction in various EFL settings worldwide, few studies have systematically investigated its impact on vocabulary learning and retention in the Iraqi EFL context. The current research addresses this gap by examining the effect of a comprehensive multimodal instructional package, incorporating written definitions, colour photographs/illustrations, audio recordings of native speaker usage, and contextual video clips, on Iraqi EFL learners' immediate vocabulary learning and their ability to retain acquired vocabulary over time. Using a quasi-experimental design, this study compares the outcomes of students who experienced vocabulary instruction through integrated visual, auditory, and textual materials with those of their counterparts taught using a traditional text-based method. The overarching goal is to investigate whether multimodal instruction leads to greater vocabulary gains and to assess the durability of those gains two weeks after instruction. By grounding this investigation in well-established cognitive theories, Dual Coding Theory, Cognitive Theory of Multimedia Learning, and the Depth Processing Hypothesis, this study aims to provide both theoretical insights and practical guidance for vocabulary instruction in resource-constrained Iraqi EFL classrooms.

2. Literature Review

2.1 Theoretical Background

Multimodal instruction is theoretically based on multiple cognitive and pedagogical theories. Core to these are dual coding theory, cognitive theory of multimedia learning, cognitive load theory, and the depth processing hypothesis.

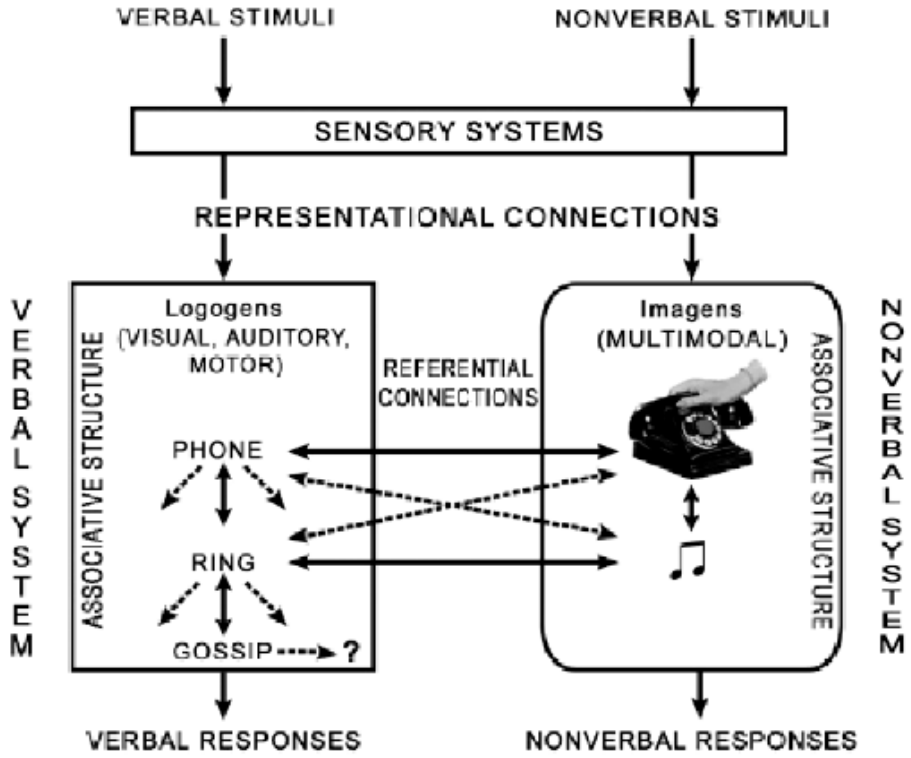
2.1.1 Dual-Coding Theory (DCT)

Allan Paivio's dual-coding theory, put forward in 1971, highlights the impact of mental imagery on memory and cognitive processes. Initially, the theory focused on memory but later expanded to include other cognitive domains (Paivio, 1971, 1986, 1991). The basic claim is that while humans can acquire knowledge through either verbal association or visual imagery, combining the two leads to better learning results (Reed, 2010). DCT suggests that the human brain processes information through two distinct channels: verbal and nonverbal systems. Sternberg (2003) referred to the nonverbal system as the imagery-based system. In the verbal system, linguistic units known as logogens are stored, while the imagery system processes picture images (imagens) (Paivio, 2014).

According to Paivio (1986) and Clark and Paivio (1991), three levels of processing are identified: representational, referential, and associative. Representational processing occurs when the stimulus

directly provokes verbal or visual representation. Referential processing establishes associations between two systems, such as a picture evoking a word or a word calling up an image. This process is crucial for developing language and creative skills (Paivio, 2014). The activation across these systems allows a child to connect a name to an object, gradually leading to mastery of structured speech (Moeser & Bregman, 1973). Associative processing refers to behavior in which an individual draws a connection between units from one system.

Figure 1. *The DCT Model of Multimodal Logogen and Imagen Units, Connections, and Implied Processes* (adopted from Kanellopoulou et al., 2019)



DCT is widely applicable to second language (L2) acquisition. Verbal and nonverbal codes operate independently, creating an additive memory retrieval effect (Paivio & Begg, 1981). When students link a new vocabulary word to both sound and image together, they remember the word better than when using a singular modality (Paivio, 1991). In her 1994 article, Rieber argued that memory task illustrations of new vocabulary work better than text only. A single memory trace may not be enough for effective learning.

Often, a picture triggers the pathway linking an image to its corresponding word (Lin, 2009), making a second memory trace more accessible. Consequently, a multimedia environment that integrates text, sound, and moving images provides an ideal setting for vocabulary acquisition (Al-Seghayer, 2001, 2016). Nonetheless, DCT suffers a limitation primarily in the "concreteness effect," where abstract ideas are difficult to visualize and therefore less likely to receive dual coding (Clark & Paivio, 1991). Concrete information is most amenable to processing in both channels, whereas abstract information is generally limited to verbal processing (Paivio, Salinas-Pazajas de la Torre, & Salim, 1994). Still, empirical test data show that DCT increases memory in many cases (Reed, 2010).

In the future (Pylyshyn, 1973), it should also examine if word and image are the sole factors at work in cognitive function or if there are other coding modalities. While Paivio (1971) originally proposed a single language-based model, Paivio and Desrochers (1980) subsequently extended the theory to encompass bilingual or multilingual contexts. In bilingual DCT, there are two separate verbal systems connected to a common imagery system. Scientists using this model have proposed several hypotheses on how the three components, autonomous, interconnected, and interdependent, function together. This study employs dual coding and bilingualism to examine the relationship between words and images, as well as the connection between a person's first language (L1) and their second language (L2). Soh (2010) outlines several key assumptions that inform these interactions within the context of language learning:

1. As in the original monolingual theory, the imagery system is independent of the two verbal systems.

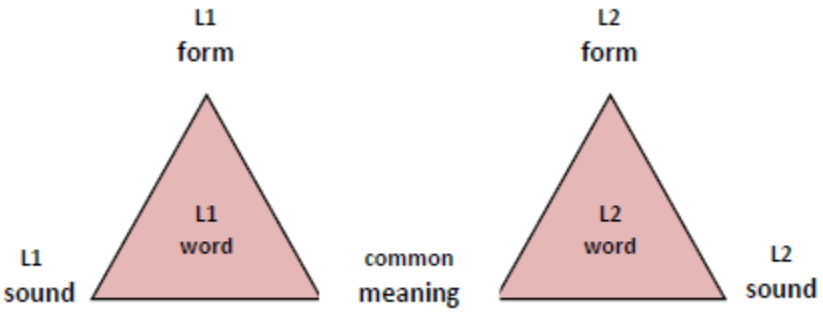
2. The three systems are linked by referential links so that imagery and verbal systems can affect one another reciprocally.

3. While the two verbal systems have referential connections to the same imagery system, they are also independent and partially related. There are two important implications for this: the extent of overlap will differ with each language pair, and the imagery system serves as a bridge allowing indirect connection between languages.

4. Bilingual people have two separate and yet interrelated verbal systems, so change in language input can lead to code-switching.

Essentially, when students pull up vocabulary from one verbal system and start recoding it in the other verbal system, practice confers on these words fresh meaning. This is a typical way to gain vocabulary with the added skill of generating translation equivalents across languages.

Figure 2. *Sharing Meanings between Two Languages through Code-Switching* (Adopted from Soh, 2010)



Despite significant advancements in this cognitive theory over the years, the bilingual DCT still requires considerable further study to assess its validity. In particular, more research is needed to understand the relationship between a learner's first language (L1) and second language (L2) and how the verbal systems of these two languages interact. A significant limitation arises when there is a lack of information in the first language (L1) compared to the knowledge in the second language (L2). For example, if a learner tries to understand an L2 concept that lacks an equivalent in L1, then there is no existing connection between the two languages. This lack of connection means that, in such cases, the learner cannot effectively apply the relevant theories. When this link exists, it can either enhance learning or cause interference.

2.1.2 Cognitive Theory of Multimedia Learning (CTML)

CTML studies how people acquire knowledge through a process of multimedia presentation (Mayer, 1997, 2001, 2005, 2014). Drawing on Wittrock's (1974) generative theory and Paivio's (1986) DCT, CTML extends the former by framing the learner as an active "knowledge constructor," not merely a passive recipient of information, who intentionally chooses and associates material, oral, and visual information between two distinct kinds of information (Mayer, 1997). This theory emphasizes that students make relationships between words and images, interpreting images and words in an active process for permanence and memory, so that learning becomes meaningful (Mayer, 1997, 2005). One

major difference between CTML and DCT is that CTML centres around an active, responsible learner in constructing information.

Mayer (2005, 2014) identifies three basic assumptions that underlie the learning process in CTML. The first is based on two cognitive pathways: the auditory/verbal channel and the visual/pictorial channel, like DCT, which postulates that sensory memory processes information through a couple of different pathways: auditory and visual. The first allows for processing auditory input (e.g., narration and spoken word) while the second processes visual input (animation, video, and graph). This type of reasoning rests on the premise that working memory (per cognitive load theory) has a limited capacity to process information in a given channel at one given time. Cognitive overload occurs when task demand exceeds this processing threshold. Mayer (2014) states that working memory is limited to maintaining a small volume of visual or verbal information simultaneously; thus, only a portion of an illustration or narration retained is retained in the relevant mental model at any given time.

Active processing: CTML describes learning as a continually evolving process in which learners choose and construct words and images, then link these with knowledge they have already retrieved from long-term memory. Mayer (2014) explains that integrating new information can involve particular cognitive functions: selecting pertinent verbal and visually available data for working memory, structuring these images into verbal and visual models, and incorporating those models into their own representation from current knowledge.

Multimedia input, particularly words and images, enters sensory memory from the environment through our senses (eyes and ears), as illustrated in Figure 3. Visual sensory memory briefly stores images and written text, while auditory sensory memory temporarily holds speech and sounds. This process occurs when learners focus on selecting particular words and images for further processing. Once this information is selected, it enters working memory, where structural relationships are established, and data is temporarily stored and manipulated.

Mayer (2014) theorizes that working memory consists of two interacting components: raw sensory information (both auditory and visual) and the information generated from it, which consists of verbal and pictorial representations that are interconnected.

Long-term memory serves as a vast storage system for information that can endure over time. Meaningful learning relies on the ongoing interaction between working memory and long-term memory (Mayer, 2005, 2014). In this interaction, relevant prior knowledge stored in long-term memory is

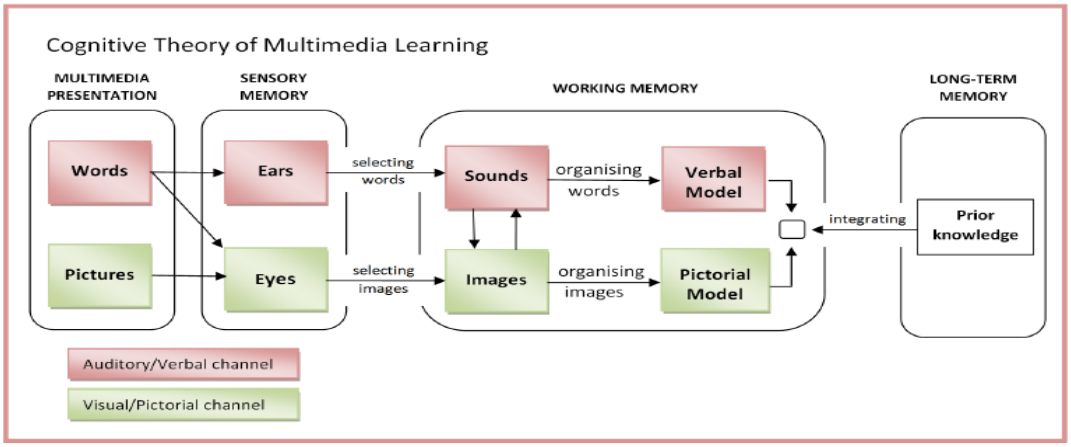
activated and connected with pertinent information entering working memory. This process aids learners in constructing logical mental representations (Mayer, 2014).

CTML follows two core principles: the multimedia principle and the temporal contiguity principle. The multimedia principle aligns with the assumptions behind dual coding theory, which posits that individuals learn more effectively when using a combination of words and images rather than relying on words alone (Mayer, 2014). This principle applies not only to written text and pictures but also to all visual elements, ranging from illustrations to video presentations (Butcher, 2006, 2014), as well as all verbal elements, including text, audio, and sound effects. By presenting information in various formats, we can increase working memory capacity and enhance the ability to recall information from long-term memory later on (Sweller, 2005).

The temporal contiguity principle, on the other hand, mitigates the risk of undue cognitive load by suggesting that deep learning occurs more frequently when corresponding verbal and visual content is presented simultaneously rather than sequentially (Mayer & Fiorella, 2014). Relationships can only be formed between closely related words and images within working memory when these elements are presented at the same time; therefore, simultaneous presentation is usually more effective than separate or consecutive communication (Mayer, 2008; Rusanganwa, 2015).

In addition to these two principles, CTML offers a more comprehensive evidence-informed guide to teaching, which Mayer categorizes based on cognitive functions (Koscianski & Zanotto, 2014). These concepts have been empirically validated (Mayer, 2008) and applied to the design of instructional materials (Herrington & Oliver, 2000; Park & Hannafin, 1993) as well as to predicting learning events (Mayer, 2001).

Figure 3. Cognitive Theory of Multimedia Learning (Adopted from Mayer, 2014, p. 52)



The first group, which is intended to reduce extraneous cognitive load, consists of the coherence effect (removing irrelevant data), signaling effect (focusing on important material to help students track attention), redundancy effect (stopping simultaneous presentation of the same text and narration on screen), spatial contiguity effect (showing similar text and image in the same spot), and temporal contiguity effect (presenting verbal and visual information together). Category two, essential processing, consists of the segmenting effect (segmenting difficult information into clear and digestible steps), pre-training effect (providing important information before the first lesson), and modality effect (using spoken rather than written text when visually presented).

The third class is designed to facilitate generative processing through a multimedia effect that combines text and image, along with a personalization approach that uses a conversational style and addresses the audience directly rather than in a formal third person. In conclusion, two broad frameworks, dual-coding theory (including bilingual dual-coding theory) and the cognitive theory of multimedia learning, reach the conclusion that multimedia conditions enhance learning by simultaneously presenting textual and visual information (Mayer, 2014; Paivio, 1986). Material that links these two modes minimizes mindless mental effort and stimulates more profound processing (Mayer & Fiorella, 2014). Recent insights have increasingly shaped contemporary studies on second language vocabulary acquisition, where watching film in the original language with subtitles is viewed as a highly effective instructional method.

2.1.3 Cognitive Load Theory (CLT)

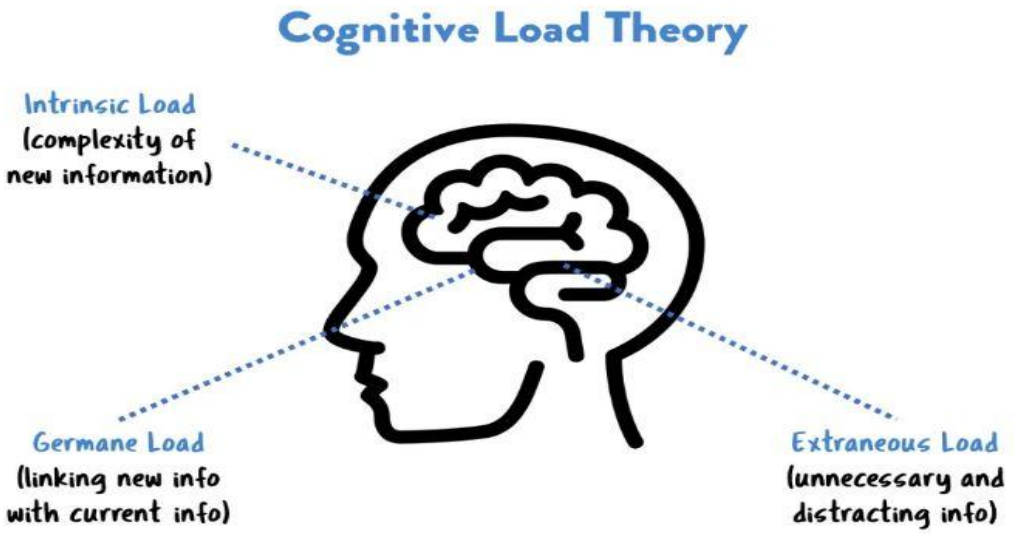
CLT was postulated by cognitive scientists and asserts that human cognitive capacity is essentially constrained. Working memory (WM) is only able to store information that can occur for a small number of items at any one time (Cowan, 2001). While WM has relatively limited capacity to consolidate simple decontextualized knowledge items, long-term memory (LTM) has unlimited capacity to store complex contextualized knowledge structures called "schema." For CLT, the main focus of instructional design is to modulate cognitive load on WM, which enables schema building in LTM (Paas, Renkl, & Sweller, 2003). This theory is particularly relevant to designing instruction using short videos, as these short films are rich in subject matter and can easily overwhelm learners' relatively limited WM capacity if not orchestrated properly.

CLT identifies three types of cognitive load that must be managed: intrinsic, extraneous, and germane. Intrinsic cognitive load refers to the inherent difficulty and complexity of the material itself. For EFL learners, the intrinsic load is relatively low for the present simple tense, as it is a grammatically

straightforward concept. In contrast, the intrinsic load increases for more complex structures, such as the passive voice or conditional sentences (Sweller, Ayres, & Kalyuga, 2011). The first type is intrinsic cognitive load, which is inherent to the content itself and cannot be reduced through design changes. The second type is extraneous cognitive load, which refers to how information is presented and the demands placed on the teaching process. When content is poorly designed, such as with unclear explanations, irrelevant images in short films, or expectations that learners must synthesize information from different sources, it creates a high level of extraneous cognitive load (Tarmizi & Sweller, 1988). This type of load wastes working memory capacity and does not contribute to effective learning.

The third variety of cognitive load, germane load, describes the constructive cognitive effort that learners put into processing information, constructing schema, and encoding this schema into LTM (Sweller et al., 2019). The goal of CLT-informed instructional design is to reduce extraneous load so WM resources can be allocated to germane processing. For example, when passive voice is covered using a short film, extraneous load can be minimized by bringing together a short explanation, a few clear examples, and a visual in a film exercise rather than spreading them around with clips or other materials or with other components. This strategy allows learners to focus cognitive load on learning the grammatical rule itself (germane load) rather than making their way through distracting or fragmented input (extraneous load).

Figure 4. *Cognitive Load Theory* (Paas et al., 2020)



Research corroborates the method: using audiovisual information through animated films, among others (Ammade & Ziska, 2021), contributes to improved clarity and reduction of unwarranted processing load. Using structured multimedia in civic and language lessons leads to improved outcomes achieved by learners as the instructional design is consistent with cognitive principles (Ardika, Agustiana, & Dibia, 2022) in both civic and language groups. This study of audiovisual media in adult education also emphasizes the reduction of extraneous stimuli for maximal learner engagement and schema formation (Nicolaou & Kalliris, 2020). Interactive pedagogical tools such as online quizzes also assist in delivering instruction without unnecessary stress and with direct focus on language skills (Oktaria, Rohmayadevi, & Murwantono, 2021). Furthermore, digital comic media can provide contextually focused input that reduces distractors and aids in germane processing (Susanti & Wibawa, 2022). Lastly, a technology-facilitated setting supported by high-level teacher self-efficacy has been shown to boost both student engagement and use of cognitive resources (Yang, 2024). Instructional grammar exercises are easier to structure and use in a strategy that is clear, direct, and within the wider cognitive architecture framework.

2.1.4 Depth Processing Hypothesis and Involvement Load Hypothesis

These modality-specific theories notwithstanding, general cognitive accounts of memory and retention support the efficacy of multimodal vocabulary instruction. Craik and Lockhart's (1972) Depth Processing Hypothesis states that retention is linked to the depth of initial processing used to process content; deeper semantic processing facilitates a stable trace of what is recalled compared to shallow perceptual processing. Multimodal tasks that require learners to link visual, auditory, and textual information stimulate deeper-level semantic thought. Similarly, Laufer and Hulstijn's (2001) Involvement Load Hypothesis posits that task-induced factors of need, search, and evaluation predict vocabulary retention. Multimodal tasks that require learners to find meaning through various methods and evaluate contextual appropriateness lead to a high involvement load that promotes retention.

2.2 Empirical Studies on Multimodal Instruction and Vocabulary Learning

The effectiveness of multimodal instruction for vocabulary learning has been extensively investigated across diverse EFL contexts, yielding a substantial body of empirical evidence that consistently supports the superiority of multimodal approaches over traditional text-only methods. This section reviews key studies that have examined various aspects of multimodal vocabulary instruction, including multimedia glossing, captioned video, annotated materials, and integrated multimedia packages.

2.2.1 Studies on Multimedia Glossing and Annotations

One of the most researched areas in multimodal vocabulary instruction is the use of multimedia glosses, annotations that provide definitions, translations, images, audio, or video to support word learning during reading or listening tasks. Chun and Plass (1996) conducted a foundational experiment investigating the effect of multimedia annotation on vocabulary learning in a computer-assisted language learning context. They found that learners who received visual and verbal annotations remembered significantly more words than those who received text-only glosses. The study demonstrated that the combination of text, pictures, and video annotations provided learners with multiple retrieval pathways, enhancing both immediate recognition and delayed recall.

Extending this line of inquiry, Plass et al. (1998) discovered that a multimedia instructional program with integrated visual and verbal explanations for unfamiliar words produced better performance on both immediate and delayed vocabulary tests compared to text-only instruction. Their findings emphasized that learners who could access both visual and verbal information were better able to construct coherent mental representations of new words, leading to more robust and durable vocabulary knowledge.

Yanguas (2009) investigated the contribution of different types of multimedia glosses, including textual glosses, pictorial glosses, and a mixture of both, on vocabulary learning and reading comprehension. The results indicated that the combined gloss condition produced the most vocabulary enhancement, consistent with the multimedia principle. Importantly, Yanguas found that pictorial glosses alone were more effective than textual glosses alone, but the combination of both yielded the highest gains, supporting the additive effect of dual coding.

More recently, Mahdi, Mohsen, and Almanea (2024) performed a second-round meta-analysis of multimedia gloss and vocabulary learning and observed a moderate overall effect size supporting multimedia over single-mode gloss, with greater benefit found when glosses included textual and pictorial input. The meta-analysis synthesized findings from numerous studies across various educational contexts, providing robust quantitative evidence for the effectiveness of multimedia glossing. Similarly, Ramezanali and Faez (2019) reported a meta-analytic effect between multimodal glossing and single-mode glossing in learning L2 vocabulary, which was strongest when a combination of visual and auditory modalities was used. Their analysis of 34 studies confirmed that multimedia glosses significantly outperform textual glosses alone, with effect sizes ranging from moderate to large depending on the modality combination.

In the Iraqi context specifically, Al-Seghayer (2001, 2016) conducted research that has been influential in shaping understanding of multimedia effects in vocabulary learning. While his studies were primarily situated in other contexts, his findings on the effectiveness of video annotations over still pictures or text-only annotations have direct relevance to Iraqi EFL settings. Al-Seghayer (2001) found that video clips, which provide dynamic contextual information and authentic language use, were significantly more effective than still images for vocabulary acquisition, as they conveyed temporal and motion information that static images could not capture.

2.2.2 Studies on Captioned and Subtitled Video

Beyond glossing, multiple multimodal teaching paradigms have been studied. According to Sydorenko (2010), students learned new vocabulary from both captioned and non-captioned videos, and combining audio, video, and text captions led to improvement in form recognition and meaning recall. Her research demonstrated that captioned video provides learners with simultaneous access to auditory input, visual context, and textual support, creating optimal conditions for vocabulary acquisition. Learners in the captioned video condition outperformed those who watched videos without captions on both form recognition and meaning recall measures.

Mohsen and Balakumar (2011) reviewed multimedia vocabulary learning studies and discovered that combining auditory and visual input is very beneficial in the vocabulary learning process. Their comprehensive review synthesized findings from over 50 empirical studies, concluding that the integration of audio and visual modalities consistently produced larger vocabulary gains than single-modality instruction. The review emphasized that the effectiveness of multimodal instruction depends on the quality and relevance of the materials, as well as the cognitive engagement of the learners.

Vanderplank (2016) conducted an extensive review of research on captioned and subtitled video in language learning, finding substantial evidence that captioned video improves vocabulary acquisition, listening comprehension, and overall language proficiency. His analysis highlighted that captions provide learners with a written representation of spoken language, facilitating form-meaning connections and supporting the development of both receptive and productive vocabulary knowledge. Vanderplank emphasized that the benefits of captioned video are particularly pronounced for intermediate learners, who can attend to both visual and auditory input while using captions as a support mechanism.

Further supporting these findings, Winke et al. (2010) investigated the effects of captioning on vocabulary learning and listening comprehension. Their study found that learners who viewed videos

with captions demonstrated significantly greater vocabulary gains than those who viewed videos without captions. The researchers also noted that the effectiveness of captioning varied depending on learner proficiency levels, with intermediate learners benefiting most from the combination of captions and audio.

2.2.3 Studies on Visual Aids, Pictures, and Illustrations

The role of visual aids in vocabulary learning has been extensively documented. Lin (2009) examined the effects of animated pictures on learning action verbs, finding that animation provided dynamic visual information that helped learners understand the meaning and usage of verbs more effectively than static images or text alone. The study demonstrated that animation reduced the cognitive effort required to infer meaning from context, allowing learners to allocate more cognitive resources to encoding and consolidation.

Al-Mansour and Al-Shorman (2012) investigated the effect of computer-assisted instruction on vocabulary learning among Saudi EFL learners, finding that learners who received instruction with visual aids significantly outperformed those who received traditional text-based instruction. Their study provided evidence that computer-assisted visual instruction facilitates vocabulary acquisition by presenting words in meaningful contexts with visual support.

Gass and Selinker (2008) discussed the role of visual input in second language acquisition, emphasizing that visual aids help learners create mental images that support vocabulary retention. They argued that pictures and illustrations provide learners with concrete referents for abstract words, facilitating the establishment of form-meaning connections. This is particularly important for learners in EFL contexts where exposure to authentic language is limited.

In a more recent study, Khoshsima et al. (2017) examined the effect of visual aids on vocabulary retention among Iranian EFL learners. Their findings revealed that learners who received vocabulary instruction with visual aids demonstrated significantly better retention than those who received text-only instruction. The study also found that the combination of visual aids with textual definitions produced the best results, supporting the dual coding principle.

2.2.4 Studies on Audio and Phonological Support

The auditory modality plays a crucial role in vocabulary learning, particularly in EFL contexts where learners have limited exposure to spoken English. Erfanian et al. (2016) investigated the effect of audio-supported vocabulary instruction on Iranian EFL learners' vocabulary retention. Their study found that learners who received audio input along with written definitions and example sentences significantly

outperformed those who received text-only instruction. The audio input provided learners with accurate pronunciation models and facilitated the development of phonological representations of new words, supporting both recognition and recall.

Rubin (1995) explored the role of audio in vocabulary acquisition, emphasizing that listening to words in context helps learners develop phonological awareness and facilitates form-meaning connections. Her research demonstrated that audio input, particularly when combined with visual support, provides learners with multiple opportunities to process and consolidate new lexical items.

In the Iraqi context, Hameed (2016) examined the effect of audio-visual materials on vocabulary learning among Iraqi EFL learners. The study found that learners who received instruction with audio-visual materials demonstrated significantly better vocabulary acquisition than those who received traditional text-based instruction. The researcher attributed these findings to the combined benefits of auditory and visual input, which provided learners with more comprehensive and meaningful exposure to new words.

2.2.5 Studies on Integrated Multimodal Packages

Recent research has increasingly focused on the effectiveness of integrated multimodal packages that combine multiple modalities, text, images, audio, and video, in a single instructional intervention. Rahimi and Allahyari (2016) investigated the effect of a multimedia vocabulary program on Iranian EFL learners' vocabulary knowledge and retention. The program incorporated written definitions, example sentences, audio pronunciations, images, and video clips for each target word. The results revealed that learners who used the multimedia program significantly outperformed those who received traditional vocabulary instruction on both immediate and delayed vocabulary tests.

Similarly, Khezrlou et al. (2017) examined the effects of computer-assisted language learning (CALL) on vocabulary acquisition and retention. Their study found that learners who received multimodal instruction through CALL programs demonstrated greater vocabulary gains and better retention than learners who received traditional instruction. The researchers attributed these findings to the multiple processing opportunities provided by multimodal input.

In a meta-analytic study, Zhang and Zou (2022) synthesized findings from 48 empirical studies on multimedia-assisted vocabulary learning, concluding that multimedia instruction generally produces large effects on vocabulary acquisition and moderate effects on vocabulary retention. Their analysis identified several factors that moderate the effectiveness of multimedia instruction, including the type of multimedia materials used, the proficiency level of the learners, and the duration of the intervention.

Notably, the meta-analysis found that multimedia instruction is particularly effective for intermediate learners, as they have sufficient foundational knowledge to benefit from the additional processing opportunities provided by multimodal input.

Lee and Mayer (2015) examined the effect of video-based multimedia instruction on vocabulary learning, comparing different presentation formats. Their study found that learners who received instruction with short video clips incorporating narration, visuals, and captions demonstrated significantly better vocabulary learning than those who received instruction with still images or text alone. The researchers emphasized the importance of temporal contiguity, presenting corresponding verbal and visual information simultaneously, in facilitating vocabulary learning.

Despite this abundant evidence from various EFL contexts, empirical research on multimodal vocabulary instruction remains notably sparse in the Iraqi EFL context. While studies conducted in other settings provide valuable insights, the unique contextual challenges faced by Iraqi learners, including limited authentic English exposure, resource constraints, large class sizes, and a predominantly examination-driven pedagogical culture, necessitate localized research to determine the effectiveness of multimodal approaches in this specific context. The present study addresses this gap by empirically investigating the effects of a comprehensive multimodal instructional package on vocabulary learning and retention among intermediate Iraqi EFL learners.

Although this is based on Western or East Asian EFL settings, this has further implications since Iraq also has very limited exposure to English and scarce resources, indicating an urgent need for instruction that maximizes the effectiveness of existing materials.

Though multimodal vocabulary instruction has ample theoretical and empirical support across the globe, it continues to receive very little scholarly attention in the Iraqi EFL context. The Iraqi system of learning is still largely based on a traditional teacher-oriented approach and, for the time being, focuses on translating grammar, series memorisation tasks, and language instruction mainly in Arabic (Saeed, 2015). Public schools lack resources like libraries, audio-visual tools, and additional materials to engage learners more effectively (Saeed, 2015), while the classroom experience is characterized by large class sizes, noise pollution, and a lack of infrastructure. These challenges are highlighted by the security situation being characterized as unstable, resulting in additional interruptions to the educational process and decreased student motivation to attend school (Saeed, 2015). This often leaves text-centered vocabulary learning ineffectual without providing learners with the rich contextualised language input needed for meaningful processing and long-term retention (Al-Marzouqi, 2020).

The present study is distinguished by its comprehensive approach to multimodal vocabulary instruction, simultaneously investigating the effects of four distinct modalities, written definitions, visual illustrations, audio recordings, and video clips, on both immediate vocabulary learning and delayed retention. While previous studies have typically examined individual modalities or limited combinations, this study provides a more holistic investigation of how integrated multimodal input affects vocabulary acquisition and retention in a resource-limited Iraqi EFL context. By drawing upon dual coding theory, cognitive theory of multimedia learning, and depth processing hypothesis, this research offers both theoretical validation and practical guidance for implementing multimodal vocabulary instruction in Iraqi classrooms.

Although research in other EFL settings has found that multimodal instruction can overcome a lack of L2 exposure and extend vocabulary acquisition by multiple sensory means (Chun & Plass, 1996; Yanguas, 2009; Mahdi et al., 2024), no empirical research has yet been conducted on the effect of a multimodal-oriented teaching programme on vocabulary learning and retention of Iraqi EFL learners to date. In view of the special social and contextual problems endured by Iraqi learners and the role of multimodal pedagogy in overcoming these difficulties, there is evident necessity for conducting a study in the field that addresses these points. Accordingly, the present study aimed to address the following research questions:

1. Does using multimodal instruction improve Iraqi EFL learners' immediate vocabulary learning compared to traditional instruction?

2. Does multimodal instruction significantly increase Iraqi EFL learners' vocabulary retention over two weeks compared to traditional instruction?

By addressing these questions, the study contributes to the limited literature by extending the understanding of technology-based and multisensory instruction within the Iraqi EFL context. These findings will help provide curriculum developers, teacher trainers, and classroom practitioners with practical guidance on effective multimodal vocabulary instruction that is sensitive to the resource constraints characteristic of Iraqi educational settings. Likewise, this investigation applies both well-established cognitive and multimedia learning theories to a relatively underexplored population, further enriching theoretical discussion on the modality effect in L2 vocabulary learning.

3. Methodology

3.1 Research Design

The study employs a quasi-experimental pre-test-post-test control group design to investigate the effects of multimodal instruction on vocabulary learning and retention. The independent variable is the type of vocabulary instruction, multimodal versus traditional, and the dependent variable is vocabulary learning measured by scores on an immediate post-test and vocabulary retention measured by scores on a delayed post-test administered two weeks after the intervention. This study consists of two groups: an experimental group receives vocabulary instruction through multiple modalities (written definitions, visual illustrations, audio recordings, and video clips), and a control group receives traditional text-based instruction. Both groups complete a pre-test, an immediate post-test after the six-week intervention, and a delayed post-test two weeks later.

The rationale for using multiple modalities in the experimental group is grounded in dual coding theory (Paivio, 1971) and cognitive theory of multimedia learning (Mayer, 2014), which predict that presenting vocabulary through multiple sensory channels creates redundant memory traces and facilitates deeper semantic processing, leading to better learning and retention. The use of four distinct modalities, written definitions, visual illustrations, audio recordings, and video clips, allows the study to capture the cumulative effect of multimodal input on vocabulary acquisition.

3.2 Participants

Sixty Iraqi EFL students enrolled in an intermediate-level English class at a private language school in Baghdad participated in the academic year 2024–2025. The learners were selected from an initial pool of 85 volunteers based on their performance on the Oxford Quick Placement Test (OQPT; Allan, 2004). Only participants scoring between 40 and 47, corresponding to level B1 (intermediate), were included in the study to ensure homogeneity. The final sample consisted of 32 male and 28 female participants aged 18–25 years. Participants were randomly assigned to the experimental group (n = 30) and the control group (n = 30). All individuals spoke Arabic as their first language and had at least seven years of English education in school. Data were collected from all participants after a comprehensive consent process.

3.3 Research Instruments

Three instruments were used in the research: the Oxford Quick Placement Test, a vocabulary test, and multimodal instructional materials.

The Oxford Quick Placement Test (OQPT; Allan, 2004) was used to assess general English proficiency and ensure group conformity among students. The test contains 60 items measuring grammar, vocabulary, and reading comprehension with a reported reliability of 0.90. The OQPT was selected because it is a well-established placement test with demonstrated validity and reliability in various EFL contexts.

The researcher developed a 30-item vocabulary test to assess knowledge of 30 target words from the intermediate-level syllabus used by the institute. The target content words, including nouns, verbs, and adjectives, were not previously taught to the participants. The test was administered in three distinct forms: Form A (pre-test), Form B (immediate post-test), and Form C (delayed post-test), with items arranged differently to minimize test-retest effects. Each item presented a target word in a sentence followed by four multiple-choice options: one clear synonym or short definition and three distractors. Content validity was confirmed by a panel of three experienced EFL teachers, and internal consistency was calculated using a pilot study with 15 learners of comparable proficiency, resulting in Cronbach's alpha of 0.87. The three parallel forms were designed to assess the same 30 target words but with different item orders and slightly varied sentence contexts, ensuring consistency in difficulty level while reducing practice effects.

Multimodal instructional materials were specifically created for the experimental group. Each target word was presented on a digital slide incorporating four elements: (a) a written definition in English, (b) a high-quality color photograph or illustration depicting the word's meaning, (c) an audio recording of a native speaker using the word in a sentence, and (d) a short video clip showing the word in context. These four modalities were intentionally selected to provide comprehensive sensory input that engages both verbal and non-verbal processing systems. Each learner received the materials on a tablet computer. Conversely, the control group received the same target words through paper materials, including only a written definition, an Arabic translation, and a sample sentence. This design allows for clear comparison between the multimodal package and traditional text-only instruction.

3.4 Data Collection Procedure

The study was conducted over eight weeks. In Week 1, volunteers were randomly assigned into groups, with 60 homogeneous intermediate learners identified after the OQPT was administered. The vocabulary pre-test (Form A) was given to both groups to assess prior vocabulary knowledge in Week 2. No significant difference in pre-test scores was found between the groups (see Section 4.1). The intervention began in Week 3 and was delivered over six weeks with sessions twice a week, each lasting

45 minutes (12 sessions total). Five target words were introduced in each session and then recycled in subsequent sessions.

The experimental group experienced multimodal instruction where each word was presented using a digital slide containing a written definition, a visual illustration, an audio recording of a native speaker using the word in a sentence, and a contextual video clip. This was followed by guided activities in which students engaged in small-group work, producing their own sentences linked to the word and image and discussing the video context. The multimodal presentation was designed so that learners could access all four modalities simultaneously, allowing them to process the information through both verbal and visual channels as predicted by dual coding theory (Paivio, 1971).

The control group received standard instruction: the teacher wrote each word on the whiteboard, introduced the Arabic translation and written definition, and required students to copy the information and complete written tasks, including gap-filling and translation drills. All 30 target words were covered by both groups during the sessions. To control for teacher effect, the same instructor taught all sessions.

In Week 9, one day following the last session, the immediate vocabulary post-test (Form B) was given to both groups. In Week 11, the delayed vocabulary post-test (Form C) was given to assess retention. Both post-tests were conducted in a controlled classroom without prior notice to the students to minimize preparation effects and capture natural learning outcomes.

The decision to use a two-week retention interval was guided by previous research on vocabulary retention (e.g., Yanguas, 2009; Ramezanali & Faez, 2019), which typically employs intervals ranging from one to four weeks to assess medium-term retention. The two-week interval was considered sufficient to distinguish between short-term memory effects and more durable vocabulary knowledge, while still being manageable within the study timeline. This interval also aligns with the practical needs of educational settings, where teachers need to know whether vocabulary gains are sustained beyond the immediate learning period.

3.5 Method of Data Analysis

Quantitative analysis was performed using SPSS software (version 27). Descriptive statistics (means and standard deviations) were calculated for pre-test, immediate post-test, and delayed post-test scores. The assumption of normality was evaluated using the Kolmogorov-Smirnov test. One-way analysis of covariance (ANCOVA) was conducted to compare group performance on the immediate and delayed post-tests while controlling for any pre-existing differences, with pre-test scores used as a covariate. Pre-

test scores were compared using an independent-samples t-test. Significance was evaluated at $\alpha = .05$. Effect sizes were computed based on partial eta squared (η^2□) for ANCOVA and Cohen's d for t-tests.

The use of ANCOVA was considered appropriate for this study because it controls for pre-existing differences in vocabulary knowledge, increasing the statistical power and precision of the analysis. By using pre-test scores as a covariate, the ANCOVA adjusts for any initial differences between groups, providing a more accurate estimate of the treatment effect. The partial eta squared (η^2□) effect sizes were interpreted according to Cohen's (1988) conventions: .01 = small, .06 = moderate, and .14 = large.

4. Results

4.1 Preliminary Analysis

Data were screened for normality and baseline equivalence before answering the research questions. The Kolmogorov-Smirnov test revealed that both pre-test, immediate post-test, and delayed post-test scores were normally distributed (all $p > .05$) for both groups. An independent samples t-test was performed to ensure that the difference between the experimental group ($M = 11.27$, $SD = 2.45$) and the control group ($M = 11.33$, $SD = 2.51$) was not statistically significant at the study start, $t(58) = 0.11$, $p = .915$. Therefore, prior vocabulary knowledge was relatively similar between the two groups.

4.2 Results for Research Question One

Research Question 1: Does multimodal instruction significantly improve immediate vocabulary learning for Iraqi EFL learners compared to traditional instruction? To address this question, a one-way ANCOVA was applied to immediate post-test scores with pre-test scores used as a covariate. Descriptive statistics for the immediate post-test are shown in Table 1.

Table 1. Descriptive Statistics for Immediate Vocabulary Post-Test

Group	N	Mean	SD	Adjusted Mean
Experimental	30	23.40	3.10	23.42
Control	30	18.73	3.45	18.71

The experimental group (Table 1) obtained a higher mean score on the immediate post-test (23.40) than the control group (18.73). The adjusted difference after controlling for pre-test scores remained significant. Table 2 summarizes the ANCOVA results.

Table 2. One-Way ANCOVA for Immediate Vocabulary Post-Test Scores

Source	SS	df	MS	F	p	η^2□
Pre-test (Covariate)	88.12	1	88.12	10.45	.002	.152

Group	342.25	1	342.25	40.62	< .001	.413
Error	488.15	57	8.56			

The analysis demonstrated a statistically significant main effect of group, $F(1,57) = 40.62$, $p < .001$, with a very large effect size (η^2 = .413). The covariate (pre-test score) was also significant, $F(1,57) = 10.45$, $p = .002$. The adjusted mean of 23.42 for the experimental group and 18.71 for the control group indicates that after controlling for prior vocabulary knowledge, the multimodal instruction group achieved significantly better results than the traditional instruction group on the immediate vocabulary test. These findings suggest that multimodal instruction produced more intensive and immediate vocabulary learning.

4.3 Results for Research Question Two

Research Question 2: Does learning vocabulary via multimodal instruction increase vocabulary retention for Iraqi EFL students over two weeks compared to using standard instruction? One-way ANCOVA was also run on delayed post-test scores, taking pre-test scores as a covariate. Descriptive statistics are shown in Table 3.

Table 3. Descriptive Statistics for Delayed Vocabulary Post-Test

Group	N	Mean	SD	Adjusted Mean
Experimental	30	21.80	3.55	21.82
Control	30	14.87	3.90	14.85

The experimental group had a higher mean score on the delayed post-test (21.80) than the control group (14.87). ANCOVA results are shown in Table 4.

Table 4. One-Way ANCOVA for Delayed Vocabulary Post-Test Results

Source	SS	df	MS	F	p	η^2
Pre-test (Covariate)	102.41	1	102.41	8.96	.004	.134
Group	712.33	1	712.33	62.31	< .001	.521
Error	652.09	57	11.44			

A significant main effect of group was found, $F(1,57) = 62.31$, $p < .001$, with a very large effect size (η^2 = .521). The covariate was also significant, $F(1,57) = 8.96$, $p = .004$. The adjusted mean of 21.82 for the experimental group and 14.85 for the control group suggests that the multimodal group retained significantly more vocabulary than the control group after two weeks. These results indicate that multimodal instruction supported both initial learning and more durable longer-term retention.

Table 5. Score Changes from Immediate to Delayed Post-Test

Group	Immediate Mean	Delayed Mean	Mean Decrease	Percentage Retention
Experimental	23.40	21.80	1.60	93.2%
Control	18.73	14.87	3.86	79.4%

As shown in Table 5, the experimental group maintained 93.2% of their vocabulary knowledge over the two-week interval, compared to only 79.4% retention in the control group. This substantial difference in retention rates further demonstrates the durability of vocabulary learning achieved through multimodal instruction.

5. Discussion

This section presents our quantitative findings and their significance according to the theoretical framework and previous studies covered in the literature review. The research questions are restated, and results are discussed thematically, with interpretation informed by theoretical and empirical perspectives.

5.1 Discussion of Research Question One

The first research question addressed whether multimodal instruction enhances vocabulary learning in the short term compared to traditional teaching methods. Findings showed that the multimodal group significantly outperformed the traditional group on the immediate post-test with a large effect size (η^2 = .413). This result is in line with the prediction of Dual Coding Theory (Paivio, 1971), in which information processed through verbal and non-verbal channels forms richer memory representations. The multimodal materials allowed learners to engage with images, audio, and video in addition to written text, enabling them to encode new lexical items more frequently and helping to build a robust semantic network for each word.

Specifically, the digital slides used in this study provided learners with simultaneous access to four distinct types of information: (a) a written definition in English, which engaged the verbal processing system; (b) a high-quality color photograph or illustration, which engaged the visual processing system; (c) an audio recording of a native speaker using the word in a sentence, which provided phonological input and engaged the auditory processing system; and (d) a short video clip showing the word in context, which combined visual, auditory, and contextual information. This integration of multiple modalities created multiple memory traces for each vocabulary item. When learners attempted to recall a word later, they could access any of these traces, increasing the likelihood of successful retrieval. This

aligns with Paivio's (1986) assertion that dual coding, processing information through both verbal and non-verbal systems, produces an additive memory effect.

The effectiveness of each modality in the multimodal package can be understood through the lens of Mayer's (2014) multimedia principle, which suggests that individuals learn more effectively from words and pictures than from words alone. The visual illustrations and video clips in the experimental condition provided concrete referents for abstract concepts, making the meanings of new words more accessible and memorable. The audio recordings provided accurate pronunciation models and helped learners establish phonological representations of new words, which is particularly important in EFL contexts where learners have limited exposure to spoken English (Erfanian Mohammadi et al., 2016).

The video clips, which showed the target words in authentic or semi-authentic contexts, were especially valuable because they provided rich contextual information that helped learners understand not only the meaning but also the appropriate usage of each word. Research by Al-Seghayer (2001) has shown that video annotations are more effective than still images for vocabulary learning because they convey temporal and motion information that static images cannot capture. The video clips in this study also provided examples of how the words are used in natural discourse, supporting the development of collocational and grammatical knowledge.

The active processing assumption of Mayer's (2014) theory was also upheld, as learners chose appropriate visual and auditory information from the slides, organized it into coherent mental structures, and integrated it with their existing knowledge during small-group activities. The experimental group's activities, discussing video contexts, producing their own sentences, and connecting words to images, required deep engagement with the material, consistent with the generative processing principle of CTML. These tasks stimulated deeper semantic processing, which Craik and Lockhart's (1972) Depth Processing Hypothesis identifies as a vital component for establishing long-term memory.

The involvement load created by the multimodal tasks also contributed to the strong learning outcomes. According to Laufer and Hulstijn's (2001) Involvement Load Hypothesis, tasks that require learners to search for meaning and evaluate contextual appropriateness lead to higher involvement load and better retention. In the experimental group, learners were asked to find meaning in images, compare the word across different modalities, and evaluate the appropriateness of word meanings in various contexts. These processes required greater semantic processing and created more elaborate memory traces, supporting both immediate learning and longer-term retention.

Findings also concur with Yanguas (2009), who observed the highest vocabulary gains using a combination of textual and pictorial glosses. The present study builds on this outcome using a wider multimodal design with audio and video alongside static imagery. The large effect size found in this study (η^2 = .413) indicates that the benefits of multimodal instruction are substantial even in a resource-limited EFL context like Iraq. This finding is consistent with meta-analytic findings by Mahdi et al. (2024), who reported moderate to large effects for multimedia glosses on vocabulary learning, and Ramezanali and Faez (2019), who found that multimodal glossing produced stronger effects than single-mode glossing.

The pedagogical implications of these findings are significant. In the Iraqi EFL context, where learners have limited exposure to English outside the classroom and resources are constrained, multimodal instruction offers a viable and effective approach to vocabulary teaching. The use of digital slides incorporating multiple modalities, even in resource-limited settings, can be implemented using basic technology such as tablets, smartphones, or even projectors. The key is to ensure that the multimodal materials are well-designed and aligned with cognitive principles of learning.

5.2 Discussion of Research Question Two

The second research question addressed whether multimodal instruction increases vocabulary retention over time. The multimodal group performed significantly better than the control group on the delayed post-test, with a very large effect size (η^2 = .521). The experimental group had a mean decrease of only 1.60 points from the immediate to delayed post-test, while the control group had a mean decrease of 3.86 points, indicating that multimodal instruction supported more sustained retention.

These retention benefits can be attributed to several interconnected theoretical mechanisms. First, the dual coding effect (Paivio, 1971) provided learners with multiple retrieval options. When attempting to recall a word, learners could access either a verbal or a visual representation, or both, enhancing the likelihood of successful recall despite the delay. The presence of multiple memory traces, one for each modality, created redundancy that protected against forgetting. This is consistent with Paivio's (1986) assertion that dual coding creates "redundant memory traces" that are more resistant to decay over time.

Second, the multimodal activities promoted deeper processing (Craik & Lockhart, 1972), resulting in more elaborate and distinctive memory traces that are less prone to forgetting. The experimental group's engagement with multiple modalities, reading definitions, viewing illustrations, listening to audio, watching video clips, and discussing contexts, required deeper semantic processing than the control group's rote memorisation and translation activities. According to Craik and Lockhart (1972), deeper

processing, which involves semantic and elaborative encoding, produces more durable memory traces than shallow perceptual processing.

Third, the reduction of extraneous cognitive load and facilitation of germane processing, as suggested by Cognitive Load Theory (Sweller, 1988), enabled learners to spend a greater share of their cognitive resources constructing and preserving schemas, generating more organised and retrievable lexical knowledge. The multimodal materials were designed to reduce extraneous load by presenting information in a clear, integrated format that minimized cognitive distractions. This allowed learners to focus their cognitive resources on germane processing, constructing meaningful associations between new words and existing knowledge structures.

Fourth, the multimodal instruction provided learners with rich contextual information that supported the development of lexical networks. The video clips and illustrations provided situational and visual contexts for each word, helping learners understand how the word is used in authentic situations. This contextual information facilitated the formation of semantic connections between new words and existing knowledge, making the lexical knowledge more integrated and accessible.

Retention results are consistent with previous research. In their meta-analysis, Ramezanali and Faez (2019) found that multimodal glosses produced superior performance on delayed post-tests, providing evidence that multimodal encoding improves long-term retention. Sydorenko (2010) also demonstrated that vocabulary gains from captioned video were sustained over time. Building on this evidence, the current research shows that a multimodal package with multiple media elements, images, audio, video, and text, can provide more sustained vocabulary storage in an Iraqi EFL setting where learners lack adequate exposure to English outside the classroom and depend on instruction for lexical development.

The retention benefits of multimodal instruction are particularly significant in the Iraqi EFL context. Traditional instruction, which emphasizes rote memorisation and translation, tends to produce vocabulary knowledge that is fragile and easily forgotten. In contrast, multimodal instruction creates robust lexical representations that are more resistant to forgetting, as evidenced by the experimental group's superior retention of 93.2% of vocabulary knowledge over two weeks compared to 79.4% retention in the control group. This finding has important implications for curriculum design and teaching practice, suggesting that investment in multimodal materials may yield long-term benefits that justify the initial resource costs.

The study also contributes to theoretical understanding of vocabulary acquisition in EFL contexts. The results confirm that the principles of dual coding theory and cognitive theory of multimedia learning

apply not only to well-resourced educational settings but also to resource-constrained environments like Iraqi classrooms. The findings extend the generalizability of these theoretical frameworks to a new population and context, enriching the theoretical discussion on modality effects in L2 vocabulary learning.

6. Conclusion

This research aimed to investigate the significance of multimodal instruction in learning and retaining vocabulary among intermediate Iraqi EFL learners. The findings confirmed that learners using vocabulary instruction integrating visual, auditory, and written sources were significantly superior after instruction and two weeks post-instruction compared to traditional text-based instruction. The large effect sizes on both learning and retention demonstrate both practical and theoretical significance of the multimodal model.

6.1 Summary of Findings

The study produced three main findings:

First, the multimodal group (n = 30) significantly outperformed the control group (n = 30) on the immediate vocabulary post-test ($F(1,57) = 40.62$, $p < .001$, $\eta^2 = .413$). This demonstrates that the integrated use of written definitions, visual illustrations, audio recordings, and video clips produced significantly better immediate vocabulary learning than traditional text-based instruction.

Second, the multimodal group demonstrated significantly better vocabulary retention over a two-week period ($F(1,57) = 62.31$, $p < .001$, $\eta^2 = .521$). The experimental group maintained 93.2% of their vocabulary knowledge, compared to only 79.4% retention in the control group, indicating that multimodal instruction produces more durable lexical knowledge.

Third, the large effect sizes for both immediate learning and retention indicate that multimodal instruction has substantial practical significance in the Iraqi EFL context. These effects were consistent across both research questions, suggesting that the benefits of multimodal instruction are robust and reliable.

6.2 Pedagogical Implications

The findings of this study have several important pedagogical implications for Iraqi EFL professionals and educators in similar resource-constrained contexts:

Table 6. Pedagogical Implications for Multimodal Vocabulary Instruction in Iraqi EFL Classrooms

Modality	Instructional Application	Theoretical Basis	Implementation Strategy
Written Definitions	Provide clear, comprehensible definitions in English	Verbal processing (DCT)	Use simple language, include example sentences, avoid overly complex definitions
Visual Illustrations	Use photographs, drawings, or diagrams to represent word meanings	Visual processing (DCT)	Ensure images are high-quality, relevant, and culturally appropriate
Audio Recordings	Provide native speaker pronunciations and usage examples	Auditory processing (DCT, CTML)	Include multiple examples of words in varied contexts
Video Clips	Show words in authentic or semi-authentic contexts	Dynamic contextual learning (CTML)	Use short clips (30-60 seconds), ensure clear demonstration of word meaning and usage

For Iraqi EFL professionals, the incorporation of multimodal resources, even simple digital slides on multimedia mediums (e.g., images, audio, and brief videos), can greatly enhance vocabulary instruction. The specific findings regarding each modality suggest that:

The written definitions provide a foundation for verbal processing, but should be supplemented with visual and auditory input to create dual coding effects. The visual illustrations help learners form mental images that support memory and understanding, particularly for concrete nouns and actions. The audio recordings provide accurate pronunciation models, which are crucial for developing receptive and productive vocabulary knowledge. The video clips offer rich contextual information that helps learners understand how words are used in authentic situations, supporting the development of collocational and grammatical knowledge.

For curriculum designers and material developers, the findings suggest that integrating "multimodal" elements into textbooks and resources, especially in schools with limited exposure to real English, can significantly improve vocabulary learning and retention. This may involve creating supplementary materials that provide visual, auditory, and contextual support for vocabulary items. In settings with limited technology, even simple materials such as printed images or audio recordings played on a speaker can provide multimodal input.

For teacher training, the findings highlight the need to prepare teachers to design multimodal activities that are compatible with multimedia learning standards. Teachers should be trained to use

multimedia materials in ways that reduce rather than detract from cognitive load. This includes following Mayer's (2014) principles for effective multimedia design, such as eliminating irrelevant information, presenting corresponding verbal and visual information simultaneously, and segmenting complex information into manageable steps.

For classroom practitioners, the findings offer practical guidance on implementing multimodal vocabulary instruction. Teachers can incorporate multimodal activities into their lessons by: (a) using images and real objects to illustrate word meanings; (b) playing audio recordings of words and sentences; (c) showing short video clips that demonstrate word usage; and (d) designing tasks that engage multiple modalities simultaneously, such as watching a video with captions and writing sentences using the target vocabulary.

Furthermore, given the existing resource constraints identified in Iraqi public schools (Saeed, 2015), low-cost solutions such as using mobile devices to display multimodal content may be effective for instructional improvement. Smartphones, tablets, and even projectors can be used to present multimodal materials at minimal cost. Teachers can also create their own multimodal materials using free or low-cost digital tools, such as creating audio recordings with a smartphone or finding relevant images and videos online.

The practical significance of these findings is substantial. In a context where resources are limited and traditional instruction often fails to produce lasting vocabulary knowledge, multimodal instruction offers a viable and effective alternative. The large effect sizes observed in this study indicate that even modest investments in multimodal materials can produce significant improvements in vocabulary learning and retention.

6.3 Contributions to Knowledge

This study makes several important contributions to knowledge in the field of L2 vocabulary acquisition:

First, it provides empirical evidence for the effectiveness of multimodal vocabulary instruction in the Iraqi EFL context, filling a significant gap in the literature. While multimodal approaches have been extensively studied in other contexts, this is one of the first studies to systematically investigate their effects in Iraq.

Second, it extends the application of dual coding theory (Paivio, 1971), cognitive theory of multimedia learning (Mayer, 2014), and depth processing hypothesis (Craik & Lockhart, 1972) to a new

population and context. The findings confirm that these theoretical frameworks are applicable not only to well-resourced educational settings but also to resource-constrained environments.

Third, it provides empirical evidence for the specific contributions of different modalities, written definitions, visual illustrations, audio recordings, and video clips, to vocabulary learning and retention. This detailed analysis of modality effects offers insights that can inform the design of multimodal materials and instruction.

Fourth, it offers practical guidance for implementing multimodal vocabulary instruction in resource-constrained contexts. The study demonstrates that effective multimodal instruction does not require expensive technology or sophisticated equipment; even basic multimodal materials can produce significant learning gains.

6.4 Limitations of the Study

The study also has several limitations that should be acknowledged. First, the sample was drawn from one private language school with only intermediate-level learners, which may impact the generalizability of the findings to novice, advanced, or students in public school contexts. Future research should include learners at different proficiency levels and from various institutional settings.

Second, the intervention was relatively brief, comprising only six weeks, and the retention interval was only two weeks, so the findings do not reflect longer-term retention. Longitudinal studies with extended retention periods would provide a more complete understanding of the durability of multimodal learning effects.

Third, the multimodal materials used a particular combination of image, audio, and video; different combinations of modalities might produce different results. Future research should systematically compare different modality combinations to determine the most effective configurations for vocabulary teaching.

Fourth, the study used only receptive multiple-choice vocabulary tests, which do not measure learners' productive use of target words. Future research should include productive measures, such as writing tasks or oral production, to assess whether multimodal instruction supports productive vocabulary knowledge.

Fifth, the study did not examine individual difference variables such as learning style, working memory capacity, or motivation. Future research should investigate how these variables moderate the effects of multimodal instruction, as different learners may benefit differently from various modality combinations.

6.5 Recommendations for Future Research

Based on the findings and limitations of this study, several recommendations for future research are offered:

First, the study should be replicated with larger and more diverse samples at different proficiency levels and institutional contexts in Iraq. This would enhance the generalizability of the findings and allow for sub-group analyses to identify which learners benefit most from multimodal instruction.

Second, longitudinal studies with extended retention periods (e.g., 3-6 months) would provide a more complete understanding of the sustainability of multimodal learning effects. This is important for determining whether the retention benefits observed in this study persist over longer periods.

Third, research should systematically compare different combinations of modalities to identify the most effective configurations for vocabulary teaching. For example, video versus still images, audio versus text, and different combinations of these modalities could be compared to determine the optimal multimodal package.

Fourth, research should include productive measures of vocabulary knowledge, such as writing tasks or oral production, to assess the full range of multimodal instruction effects. This would provide a more comprehensive understanding of how multimodal instruction supports vocabulary development.

Fifth, future research should investigate the effects of individual difference variables, such as learning style, working memory capacity, and motivation, on the effectiveness of multimodal instruction. Understanding how different learners respond to multimodal instruction could inform the development of differentiated instructional approaches.

Sixth, research should explore the integration of multimodal instruction with other pedagogical approaches, such as task-based language teaching, to examine whether multimodal elements can enhance the effectiveness of these approaches. This would contribute to the development of more comprehensive and effective instructional models.

Finally, future research should investigate the cost-effectiveness of multimodal instruction in resource-constrained settings, examining the relationship between investment in multimodal materials and learning outcomes. This would provide practical guidance for educational decision-makers on how to allocate limited resources effectively.

6.6 Concluding Remarks

In conclusion, this study provides strong empirical evidence for the effectiveness of multimodal vocabulary instruction in the Iraqi EFL context. The findings confirm that integrating written

definitions, visual illustrations, audio recordings, and video clips produces significantly better vocabulary learning and retention compared to traditional text-based instruction. The large effect sizes observed indicate that multimodal instruction has substantial practical significance, offering a viable and effective approach to vocabulary teaching in resource-constrained settings.

The theoretical contributions of the study are also significant. The findings provide empirical support for dual coding theory, cognitive theory of multimedia learning, and depth processing hypothesis in a new population and context, enriching theoretical discussion on modality effects in L2 vocabulary learning. The study demonstrates that these theoretical principles are applicable across diverse educational settings, supporting the universality of cognitive learning principles.

The pedagogical implications are clear: Iraqi EFL teachers, curriculum designers, and material developers should embrace multimodal approaches to vocabulary instruction. Even with limited resources, simple multimodal materials can significantly improve vocabulary learning and retention. The key is to ensure that multimodal materials are well-designed and aligned with cognitive principles of learning.

Ultimately, this study contributes to improving EFL vocabulary instruction in Iraq by providing evidence-based guidance for implementing effective multimodal approaches. By addressing the challenges of limited authentic English exposure and resource constraints, multimodal instruction offers a pathway to more effective vocabulary teaching and learning in Iraqi educational settings.

References

Allan, D. (2004). *Oxford Placement Test 1: Test Pack*. Oxford University Press.

Al-Mansour, N. S., & Al-Shorman, R. A. (2012). The effect of computer-assisted instruction on Saudi university students' learning of English. *Journal of King Saud University - Languages and Translation, 24*(1), 51-57.

Al-Marzouqi, A. H. (2020). The impact of rote learning and translation methods on vocabulary acquisition among Iraqi EFL learners. *Journal of University of Babylon for Humanities, 28*(3), 45–57.

Al-Seghayer, K. (2001). The effect of multimedia annotation modes on L2 vocabulary acquisition: A comparative study. *Language Learning & Technology, 5*(1), 202–232.

Al-Seghayer, K. (2016). Celebrating the story of my first contribution to CALL. *Language Learning & Technology*, *20*(3), 179–184.

Ammade, S., & Ziska, I. (2021). Animation film in writing teaching instruction on narrative text: Working or not working? *Language Circle: Journal of Language and Literature, 16*(1), 81–88.

Ardika, N., Agustiana, I., & Dibia, I. (2022). Character and learning outcomes of civics in VCT learning assisted by audio visual media. *Indonesian Values and Character Education Journal, 5*(1), 19–31.

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology, 98*(1), 182–197.

Butcher, K. R. (2014). The multimedia principle. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 174–205). Cambridge University Press.

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. *The Modern Language Journal, 80*(2), 183–198.

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3*(3), 149–210.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114.

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.

Erfanian Mohammadi, J., Khalaji, H., & Naderi, F. (2016). The effect of audio-supported vocabulary instruction on Iranian EFL learners' vocabulary retention. *International Journal of English Linguistics, 6*(3), 178-188.

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course* (3rd ed.). Routledge.

Hameed, A. F. (2016). The effect of audio-visual materials on vocabulary learning among Iraqi EFL learners. *Journal of Educational Sciences, 27*(4), 45-62.

Herrington, J., & Oliver, R. (2000). An instructional design framework for authentic learning environments. *Educational Technology Research and Development, 48*(3), 23–48.

Kanellopoulou, C., Kermanidis, K., & Giannakoulopoulos, A. (2019). The dual coding theory and its applications in language learning. *International Journal of Language and Linguistics, 7*(3), 124-135.

Khoshsima, H., Rezaeian, H., & Rezaeian, M. (2017). The effect of visual aids on vocabulary retention among Iranian EFL learners. *Journal of Language Teaching and Research, 8*(4), 765-772.

Khezrlou, S., Ellis, R., & Sadeghi, K. (2017). Effects of computer-assisted language learning on vocabulary acquisition and retention. *Computer Assisted Language Learning, 30*(3-4), 283-304.

Koscianski, A., & Zanotto, D. D. C. F. (2014). A design model for educational multimedia software. *Creative Education, 5*(22), 2003–2016.

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

Lee, H., & Mayer, R. E. (2015). Video-based multimedia instruction for vocabulary learning. *Educational Technology Research and Development, 63*(5), 715-738.

Lin, C.-C. (2009). Learning action verbs with animation. *JALT CALL Journal, 5*(1), 23–40.

Mahdi, H. S., Mohsen, M. A., & Almanea, M. (2024). Multimedia glosses and second language vocabulary learning: A second-round meta-analysis. *Acta Psychologica, 248*, Article 104341.

Mayer, R. E. (1997). Multimedia learning: Are we asking the right questions? *Educational Psychologist, 32*(1), 1–19.

Mayer, R. E. (2001). *Multimedia learning*. Cambridge University Press.

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). Cambridge University Press.

Mayer, R. E. (2008). Applying the science of learning: Evidence-based principles for the design of multimedia instruction. *American Psychologist, 63*(8), 760–769.

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 43–71). Cambridge University Press.

Mayer, R. E., & Fiorella, L. (2014). Principles for reducing extraneous processing in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (2nd ed., pp. 279–315). Cambridge University Press.

Moeser, S. D., & Bregman, A. S. (1973). Imagery and language acquisition. *Journal of Verbal Learning and Verbal Behavior, 12*(1), 91–98.

Mohsen, M. A., & Balakumar, M. (2011). A review of multimedia glosses and their effects on L2 vocabulary acquisition. *English Language Teaching, 4*(1), 199–208.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P. (2006). *How vocabulary is learned*. Oxford University Press.

Nicolaou, C., & Kalliris, G. (2020). Audiovisual media communications in adult education: The case of Cyprus and Greece of adults as adult learners. *European Journal of Investigation in Health Psychology and Education, 10*(4), 967–994.

Oktaria, A., Rohmayadevi, L., & Murwantono, D. (2021). Online game quiz "Kahoot" in teaching English for students of SMP Muhammadiyah Yogyakarta. *Project: Professional Journal of English Education, 4*(2), 290–296.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: Recent developments. *Educational Psychologist, 38*(1), 1–4.

Paas, F., Kalyuga, S., & Renkl, A. (2020). *Cognitive load theory: Instructional implications for instructional design*. Routledge.

Paivio, A. (1971). *Imagery and verbal processes*. Holt, Rinehart & Winston.

Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford University Press.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology, 45*(3), 255–287.

Paivio, A. (2014). Intelligence. In *Elsevier encyclopedia of intelligence* (Vol. 47, pp. 141–158). Elsevier.

Paivio, A., & Begg, I. (1981). *Psychology of language*. Prentice-Hall.

Paivio, A., & Desrochers, A. (1980). A dual-coding approach to bilingual memory. *Canadian Journal of Psychology, 34*(4), 388–399.

Paivio, A., Walsh, M., & Bons, T. (1994). Concreteness effects on memory: When and why? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*(5), 1196–1204.

Park, I., & Hannafin, M. (1993). Empirically-based guidelines for the design of interactive multimedia. *Educational Technology Research and Development, 41*(1), 63–85.

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.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological Bulletin, 80*(1), 1–25.

Rahimi, M., & Allahyari, H. (2016). The effect of a multimedia vocabulary program on Iranian EFL learners' vocabulary knowledge and retention. *Journal of Language Teaching and Research, 7*(6), 1213-1221.

Ramezanali, M., & Faez, F. (2019). Efficacy of multimodal glossing on L2 vocabulary learning: A meta-analysis. *TESOL Quarterly, 53*(4), 1026–1057.

Reed, S. K. (2010). *Cognition: Theories and applications* (8th ed.). Wadsworth Cengage Learning.

Rieber, L. P. (1994). *Computers, graphics, and learning*. Brown and Benchmark.

Rubin, J. (1995). The contribution of audio to vocabulary acquisition. *Language Learning, 45*(4), 681-704.

Rusanganwa, J. A. (2015). Developing a multimedia instrument for technical vocabulary learning. *Computer Assisted Language Learning, 28*(1), 97–111.

Saeed, N. R. (2015). *Problems of teaching English in Iraqi public schools in Mosul for the intermediate stage from the teachers' point of view* [Unpublished master's thesis]. Middle East University, Amman, Jordan.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Soh, K. C. (2010). Bilingual dual-coding and code-switching: Implications for the L1 in L2 learning. *Journal of Linguistics and Language Teaching, 1*(2), 271–296.

Sternberg, R. J. (2003). *Cognitive psychology* (3rd ed.). Thomson/Wadsworth.

Susanti, K., & Wibawa, I. (2022). Learning socio-cultural diversity of society containing tolerance character values through digital comic media. *Mimbar PGSD Undiksha, 10*(2), 342–350.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285.

Sweller, J. (2005). Implications of cognitive load theory for multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 19–30). Cambridge University Press.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer.

Sweller, J., Ayres, P., & Kalyuga, S. (2019). *Cognitive load theory* (2nd ed.). Springer.

Sydorenko, T. (2010). Modality of input and vocabulary acquisition. *Language Learning & Technology, 14*(2), 50–73.

Tarmizi, R. A., & Sweller, J. (1988). Guidance during problem solving. *Journal of Educational Psychology, 80*(4), 424–434.

Vanderplank, R. (2016). Captioned and subtitled video in language learning. *The Modern Language Journal, 100*(S1), 204-219.

Wilkins, D. A. (1972). *Linguistics in language teaching*. Edward Arnold.

Winke, P., Gass, S. M., & Sydorenko, T. (2010). The effects of captioning on vocabulary learning and listening comprehension. *Language Learning & Technology, 14*(3), 65-89.

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist, 11*(2), 87–95.

Yang, Z. (2024). Examining the relationship between teacher self-efficacy and student engagement in technology-enhanced learning environments. *International Journal of New Developments in Education, 6*(2).

Yanguas, I. (2009). Multimedia glosses and their effect on L2 text comprehension and vocabulary learning. *Language Learning & Technology, 13*(2), 48–67.

Zhang, D., & Zou, D. (2022). Multimedia-assisted vocabulary learning: A meta-analysis. *Computer Assisted Language Learning, 35*(8), 1894-1924.

Author Information

Rusul Esmaeel Sakran (General Directorate of Education in Wasit Governorate , Ministry of Education)
goldenrose84627@gmail.com