Volume 5 | Issue 2 | 2025

http://langedutech.com

Language Education & Technology

Mapping the Multimodal Ecology of Video-Based Incidental Vocabulary Acquisition: A Systematic Review with Pedagogical Insights

Samet Çağrı Kızkapan[a] *

[a] English Language Teaching Department, Institute of Educational Sciences, Bursa Uludag University, Türkiye; 0000-0003-3761-0323

Suggested citation: Kızkapan, S.Ç. (2025). Mapping the Multimodal Ecology of Video-Based Incidental Vocabulary Acquisition: A Systematic Review with Pedagogical Insights. *Language Education and Technology (LET Journal)*, 5(2), 144-171.

Article Info

Date submitted: 27/10/2025

Date accepted: 04/02/2026

Date published: 05/02/2026

Review Article

Abstract

This study presents a PRISMA-guided systematic review synthesizing empirical research on video-based incidental English vocabulary learning in second and foreign language contexts. Using the Web of Science Core Collection as the sole database, 743 records were initially retrieved, with 103 studies retained following title–abstract screening, full-text eligibility assessment, and application of predefined inclusion criteria. The review aimed to move beyond modality comparisons by mapping the multimodal ecology through which vocabulary learning unfolds in audiovisual environments. Specifically, it examined how video-based multimodality has been theorized, which cognitive, affective, and behavioral processes mediate lexical uptake and retention, how incidental vocabulary learning has been operationalized and measured, and what pedagogical principles emerge from the accumulated evidence. The synthesis indicates that vocabulary gains from video viewing are not attributable to modality per se but to the interaction of three mechanism families: dual-channel representational enrichment, cognitive load calibration, and attention- and engagement-driven processes such as noticing, self-regulation, repetition, and task-induced elaboration. The review further demonstrates that conclusions in this domain are strongly shaped by methodological choices, particularly outcome constructs and testing timing, with recognition-based measures privileging noticing effects and delayed or productive measures foregrounding consolidation processes. Pedagogical implications highlight the need for purpose-driven captioning and glossing configurations, calibrated multimodal density, and designs that support learner control and repeated exposure. Overall, the study advances a process-oriented framework for understanding and designing video-based vocabulary learning environments.

Keywords: captioned video; incidental vocabulary acquisition; multimodal input; systematic review; video-based vocabulary learning

* Corresponding author. English Language Teaching Department, Institute of Educational Sciences, Bursa Uludag University, Bursa, Turkiye.
e-mail adress: sckizkapan@cumhuriyet.edu.tr

1. Introduction

Vocabulary knowledge constitutes a fundamental component of language proficiency, forming the foundation for learners' ability to comprehend and produce meaningful communication in a second or foreign language. As Webb and Nation (2017) highlighted, native speakers typically use between 15,000- and 20,000-word families by adulthood, whereas most L2 learners struggle to attend even 3,000–5,000, which significantly constrains their ability to comprehend authentic discourse such as films and academic texts. The number of words a learner knows is not merely an index of lexical breadth but also a determinant of their ability to acquire new vocabulary, reflecting the "Matthew effect," where those with larger vocabulary learn new words more efficiently than those with smaller ones (Webb & Nation, 2017). Vocabulary learning is inherently incremental (Teng, 2020), involving the gradual strengthening of form–meaning connections and expanding from receptive to productive mastery. Despite its centrality, achieving sufficient lexical coverage to support fluent comprehension remains an enduring challenge for L2 learners across contexts.

Given these challenges, research has sought pedagogical methods that can increase the quantity and quality of input exposure necessary for vocabulary growth. Video-based learning environments have become a promising avenue because they combine auditory, visual, and textual modalities to provide rich, contextualized input that mirrors natural language use. Rooted in Mayer's (2009) cognitive theory of multimedia learning, video-based multimodality enables learners to process linguistic information through both verbal and nonverbal channels, fostering deeper encoding of lexical items. Building upon Paivio's (2007) dual coding theory, the simultaneous presentation of spoken language, captions, and visual imagery can strengthen associative links between word form and meaning, thereby promoting more robust retention. Empirical research has confirmed that captioned and glossed videos can enhance learners' noticing of new words and comprehension of meaning (Lee & Choi, 2024; Montero Perez et al., 2015, 2018; Teng, 2020). For example, Teng (2020) found that captions helped learners form stronger form–meaning mappings, while Montero Perez et al. (2018) demonstrated that combining visual and textual enhancements supported vocabulary retention through increased attentional engagement.

Nevertheless, despite the accumulated evidence supporting video-based input, learning gains from audiovisual materials are often modest. For instance, in large-scale experiments, L2 learners typically acquired only a handful of words after hours of viewing authentic programs (e.g., Rodgers & Webb, 2020). Such limited outcomes underscore that exposure alone does not guarantee vocabulary learning; rather, the quality of multimodal processing, that is how learners attend to and integrate verbal and nonverbal information, determines the depth of lexical uptake. Boers et al. (2017a) and Boers et al. (2017b) argued that the effectiveness of multimodal annotations depends less on modality itself and more on the degree of attention that these annotations elicit. Similarly, empirical evidence from eye-tracking studies indicates that glossed captions attract additional visual attention, facilitating memory encoding through increased engagement (Godfroid et al., 2013). Thus, the pedagogical value of video-based multimodality lies not only in its multimodal richness but in its capacity to direct learners' cognitive resources effectively.

Recent studies have also highlighted theoretical ambiguities in how "multimodality" is conceptualized in vocabulary research. Boers et al. (2017a) criticized the uncritical application of dual coding theory in CALL studies, contending that the benefits of multimodal glosses are often confounded by attentional variables rather than purely by the addition of imagery. Similarly, Lee and Choi (2024) noted that although glossed captions have been theoretically linked to input enhancement and cognitive load reduction, empirical studies remain inconclusive regarding whether they stimulate active processing or merely facilitate meaning access. Furthermore, existing research has predominantly examined specific annotation types (e.g., textual vs. pictorial glosses, full vs. keyword captions) rather than exploring the broader ecology of multimodal interaction, which is the interplay among auditory, visual, and textual elements that collectively shape vocabulary learning processes. This fragmented approach has resulted in

a lack of integrative understanding of how video-based multimodal input operates as a learning mechanism.

The need for a systematic synthesis is further emphasized by methodological inconsistencies across studies. Researchers have used varying operationalizations of vocabulary learning, ranging from form recognition to meaning recall and productive use, making cross-study comparisons difficult (Peters et al., 2016; Webb & Nation, 2017). Moreover, while incidental learning has traditionally dominated this research domain (Feng & Webb, 2020; Yun, 2011), recent work suggests that intentional and incidental processes often overlap in video-based tasks, particularly when learners interact with captions and glosses (Lee & Choi, 2024; Teng, 2020). Consequently, understanding how multimodal environments mediate both incidental and deliberate learning processes requires an analytical shift from evaluating what works best to examining how learning unfolds across cognitive, affective, and behavioral dimensions.

This systematic review addresses these gaps by synthesizing the conceptual, empirical, and pedagogical dimensions of video-based vocabulary learning. Specifically, it examines how studies have theorized video-based multimodality as a mechanism for lexical development, what cognitive and affective processes mediate word learning in audiovisual contexts, how vocabulary acquisition is operationalized and measured, and what instructional principles emerge from the accumulated evidence. By mapping this multimodal ecology, the present review advances understanding from isolated modality comparisons toward an integrated model of video-based vocabulary learning. In doing so, it aims to bridge research and pedagogy, providing educators with evidence-based insights into designing effective multimodal input that not only enriches exposure but also enhances engagement, attention, and retention. Ultimately, this synthesis contributes to reorienting the field from modality-based outcomes to process-driven explanations of how multimodal environments can optimize L2 lexical growth.

In line with these gaps, the current study aimed to address the following research questions:

1. How have studies theorized video-based multimodality as a mechanism for incidental English vocabulary acquisition (e.g., noticing, involvement, cognitive load, engagement)?
2. What cognitive, affective, and behavioral processes are evidenced to mediate vocabulary uptake and retention in video-based contexts (e.g., attention allocation, depth of processing, self-regulation, interaction with controls such as pause/rewind)?
3. How is incidental learning of English vocabulary operationalized and measured across studies (task designs, immediacy vs. delayed tests, breadth vs. depth of word knowledge), and how do these choices condition the processes identified in RQ2?
4. What design principles and actionable implications for classroom/online practice emerge from the accumulated evidence (e.g., captioning configurations, glossing modalities, task sequencing, pacing/segmenting, learner control), and under what conditions are they most plausible?

2. Literature Review

2.1. Theoretical Perspectives on Video-Based Multimodality

Research on video-based vocabulary learning is grounded in cognitive theories that emphasize how learners process verbal and nonverbal information. Mayer's (1997, 2009) cognitive theory of multimedia learning and Paivio's (2007) dual coding framework posit that audiovisual input supports vocabulary learning through dual-channel processing when visual imagery and verbal information are meaningfully integrated. Empirical work on annotations and multimedia glosses consistently demonstrates that multimodal formats enhance form, meaning mapping by strengthening associative links (Akbulut, 2007; Nagata, 1999; Plass et al., 1998).

Studies on input enhancement further highlight noticing as the key mechanism. Textual enhancement and captioning have been found to direct learners' visual attention to target lexical items (Montero Perez et al., 2014, 2018; Warren et al., 2018). Eye-tracking evidence shows that captions and glosses increase fixation duration on target words, facilitating deeper processing (Montero Perez et al., 2018). Sydorenko (2010) similarly reported that captions shift attention toward orthographic representations, promoting more robust form recognition compared to audio-only conditions.

Another theoretical line concerns cognitive load. While multimodality can reduce processing difficulty by providing redundant cues, excessive visual information may overload working memory, especially for lower-proficiency learners (Hsu et al., 2013; Nation, 2013). Yet research shows that well-designed captioning and glossing configurations mitigate unnecessary load and support efficient processing of novel items (Ko, 2012, 2017). Rassaei (2018) demonstrated that augmented visual cues enhance elaboration without overwhelming learners, provided that tasks support manageable attentional allocation.

Finally, the involvement load hypothesis and depth-of-processing perspectives appear implicitly in several studies. Tasks that require higher cognitive engagement, such as inference-making or selective attention to highlighted forms, yield stronger retention (Vahedi et al., 2016; Yun, 2011). Across studies, video-based multimodality is not conceptualized merely as additional sensory channels but as a mechanism for directing and sustaining attention in ways that foster lexical encoding.

Taken together, theoretical perspectives converge on the view that multimodal video input promotes vocabulary learning through attentional, cognitive, and representational mechanisms. However, inconsistent definitions of multimodality, varying interpretations of cognitive load, and divergent uses of enhancement techniques underscore the need for systematic synthesis to clarify how these mechanisms operate across designs.

2.2. Cognitive, Affective, and Behavioral Processes in Video-Based Vocabulary Learning

Studies employing eye-tracking, think-alouds, and learner reports reveal a complex interplay of cognitive and affective processes during audiovisual learning. Attention allocation is among the most frequently documented mediators. Learners consistently direct more visual attention to captioned and glossed segments, resulting in higher lexical uptake (Montero Perez et al., 2014; Yanguas, 2009). Research on pictorial glosses shows that images further increase attentional engagement by reducing ambiguity and supporting conceptual access (Aldera & Mohsen, 2013; Al-Seghayer, 2001).

Processing depth also shapes learning outcomes. Tasks requiring integration of verbal and visual cues, such as predictive judgments (Hsieh, 2020), elaborative viewing (Fievez et al., 2023), or interactive features in glossing environments (Fakhr et al., 2021), lead to stronger retention. Learners' behavioral engagement, including using pause and replay functions, contributes to increased noticing and consolidation (Webb & Rodgers, 2009; Winke et al., 2010). Even in authentic long-form viewing, sustained engagement promotes incremental learning (Rodgers & Webb, 2020).

Affective factors additionally influence performance. Enjoyment, reduced anxiety, and perceived control have been associated with increased attention and better learning outcomes in multimedia environments (Salehi & Naserieh, 2013; Yeldham, 2018). Research on learner preference for gloss types (Çekiç, 2024) suggests that motivation interacts with cognitive effort, shaping learners' willingness to attend to enhanced cues.

Overall, evidence indicates that vocabulary gains result from coordinated cognitive, affective, and behavioral processes rather than modality alone. Yet the field lacks an integrative account of how these processes jointly operate across annotation types, viewing conditions, and learner characteristics, indicating the value of a systematic synthesis grounded in multimodal learning theory.

2.3. Operationalization and Measurement of Incidental Vocabulary Learning

Incidental vocabulary acquisition in video-based environments has been operationalized through diverse testing formats, exposure conditions, and timeframes. Most studies adopt pretest–posttest or pretest–posttest–delayed designs to assess form recognition, meaning recall, or productive use (Feng & Webb, 2020; Sydorenko, 2010; Webb & Nation, 2017). Breadth-focused measures such as form recognition tests are sensitive to initial learning, whereas meaning recall and productive tasks index deeper acquisition but often capture smaller gains (Rodgers & Webb, 2020; Webb & Rodgers, 2009).

Variation in operationalization reflects distinct assumptions about incidental learning. For example, studies on textual enhancement emphasize immediate noticing, whereas glossing studies emphasize semantic elaboration and long-term retention (Plass et al., 1998; Vahedi et al., 2016). Tasks also differ in the degree to which they permit learner control. Research that incorporates pause/rewind features (Warren et al., 2018; Winke et al., 2010) treats such behaviors as part of incidental engagement, while studies using continuous viewing treat incidental learning as fully passive (Ko, 2012; Montero Perez et al., 2014).

The heterogeneity of testing instruments and viewing conditions restricts comparability across studies. In particular, the differential use of immediate versus delayed tests complicates interpretation of retention, and the reliance on form-focused measures limits insights into depth of lexical development. These inconsistencies signal the need for a systematic review that delineates how methodological choices shape observed learning processes and outcomes.

2.4. Pedagogical Principles Emerging from Video-Based Vocabulary Research

Across the evidence base, several design principles emerge for the pedagogical use of video-based multimodality. First, captioning consistently supports lexical recognition and meaning access, especially for lower-proficiency learners (Ko, 2012, 2017; Rassaei, 2018; Winke et al., 2010). Enhanced captions that highlight target items further increase noticing, though some studies caution that excessive enhancement may increase cognitive load (Warren et al., 2018).

Second, glossing, textual, pictorial, and multimodal, enhances depth of processing and retention. Pictorial glosses yield particular benefits due to reduced semantic ambiguity and stronger dual coding (Akbulut, 2007; Al-Seghayer, 2001). However, glosses are most effective when they support rather than interrupt viewing, suggesting the need for integration rather than overlay.

Third, learner control features such as pause and replay tools facilitate behavioral engagement that contributes to deeper vocabulary processing (Winke et al., 2010). Long-form viewing research demonstrates that sustained exposure promotes incremental gains, particularly when supported by captions (Rodgers & Webb, 2020).

Finally, studies highlight the importance of aligning multimodal design with task difficulty, proficiency level, and attentional demands. Overly dense visual information can overwhelm learners (Hsu et al., 2013), while insufficient support reduces noticing (Nation, 2013).

Overall, pedagogical implications converge on the need for balanced multimodal design that fosters attention, supports manageable processing, and encourages active engagement. A systematic synthesis is warranted to refine these principles into coherent, evidence-based guidelines for instructional design across diverse learning environments.

3. Methodology

3.1. Research Design

The study employed a systematic review design guided by the principles outlined in the PRISMA 2020 statement, which provides updated, transparent reporting standards for identifying, selecting, appraising, and synthesizing studies (Page et al., 2021). As Gough et al. (2017) put forward, systematic reviews are

treated as a form of secondary research requiring explicit, rigorous, and reproducible methods for locating and evaluating published studies on a specific subject. This design ensures a structured approach to mapping the conceptual, methodological, and pedagogical dimensions of a given subject. Following this guideline, the current study investigated video-based English vocabulary learning.

3.2. Search Procedure

The search strategy was developed to locate peer-reviewed studies examining English vocabulary learning through video-based, audiovisual, or multimodal input. Consistent with PRISMA 2020's emphasis on transparency in database selection and reproducibility of search strategies (Page et al., 2021), the Web of Science Core Collection (WoS) was used as the sole database. WoS was selected because of its stable indexing practices, controlled metadata structure, and strong representation of high impact applied linguistics, psycholinguistics, and CALL journals. Restricting the search to WoS ensured a coherent, replicable dataset appropriate for a focused evidence synthesis.

The decision to rely exclusively on the Web of Science Core Collection (WoS) was theoretically and methodologically aligned with the primary aim of the study, namely to conduct a structured ecological mapping of research trends, conceptual foci, and methodological patterns in English vocabulary learning through video-based and multimodal input, rather than to exhaustively retrieve all extant studies. WoS was selected because it offers a highly curated and stable citation index with rigorous journal inclusion criteria, standardized metadata, and long-term indexing consistency, all of which are essential for ensuring internal coherence and analytical comparability in synthesis-oriented reviews.

While multi-database searches are recommended for reviews prioritizing maximum recall, such breadth often introduces substantial redundancy, inconsistent indexing practices, and metadata noise that can obscure higher-level pattern detection. In contrast, WoS provides a controlled corpus that reliably captures the core, high-impact outlets in applied linguistics, psycholinguistics, and computer-assisted language learning, the primary disciplinary homes of research on multimodal vocabulary learning. Empirical comparisons in prior synthesis research have shown that WoS alone retrieves the vast majority of highly cited and theoretically influential studies in these fields, with marginal conceptual gains from additional databases once core journals are covered.

Importantly, the objective of this review was not to estimate effect sizes or to make claims of statistical completeness, but to map dominant research trajectories, instructional modalities, and learning outcomes within a bounded and replicable scholarly ecosystem. From this perspective, restricting the search to WoS reduced database-driven heterogeneity and enhanced reproducibility, transparency, and interpretive precision. Nevertheless, the authors acknowledge that studies published in regionally indexed or practice-oriented journals may not be fully represented, and this delimitation is treated as a conscious scoping decision rather than a methodological shortcoming.

To align the search precisely with the review's scope, the Boolean string was constructed around three clusters: (a) video-based or multimodal input; (b) vocabulary learning; and (c) English as the target language. Using topic search option, the operationalized search string was:

("video-based" OR "educational video" OR "authentic video" OR "caption*" OR "subtitl*" OR "gloss*" OR "glossed video" OR "captioned video" OR "subtitled video" OR "audiovisual") AND (vocab* OR lexic* OR "word learning" OR "vocabulary knowledge" OR "vocabulary acquisition") AND (English OR ESL OR EFL OR "English vocabulary" OR "L2")

This formulation ensured that only studies reporting English vocabulary learning outcomes were retrieved, excluding research on other target languages. The search covered all years indexed in WoS up

to the search date and included only peer-reviewed journal articles. This search yielded 743 in WoS database.

3.3. Eligibility Criteria

Eligibility criteria were determined in advance and applied consistently across all screening stages, following the structured decision-making principles described by Gough et al. (2017).

3.3.1. Inclusion criteria:

1. Empirical studies examining English vocabulary learning (ESL/EFL contexts).

2. Studies using video-based, audiovisual, or multimodal input as the instructional or exposure medium.

3. Studies reporting measurable English vocabulary outcomes (e.g., form recognition, meaning recall, productive knowledge, retention).

4. Peer-reviewed journal articles.

5. Any study design (experimental, quasi-experimental, mixed methods, or qualitative) provided vocabulary learning was analyzable.

3.3.2. Exclusion criteria:

1. Studies where the target vocabulary was not English (e.g., L2 Spanish, L2 Japanese).

2. Studies using English video input but not measuring English vocabulary learning.

3. Studies focusing exclusively on reading-based or audio-only input.

4. Conceptual or theoretical papers without empirical data.

5. Dissertations, book chapters, conference proceedings, and non–peer-reviewed sources.

These refinements ensured that the corpus strictly reflected research on English vocabulary acquisition through video-based multimodal input.

3.4. Study Selection and Data Extraction

Study selection occurred in three stages: title screening, abstract screening, and full-text review. The eligibility criteria were applied at each stage to ensure consistency and transparency. A structured data extraction template captured the following elements: theoretical framing of multimodality, multimodal features investigated (e.g., captions, pictorial glosses, keyword captions), research design, participant characteristics, vocabulary measures (construct, type, timing), mediating processes (cognitive, affective, behavioral), and pedagogical implications. This structure aligns with PRISMA 2020 expectations for transparent reporting of extracted variables (Page et al., 2021).

This search yielded 743 records in the Web of Science database. An initial screening based on titles, abstracts, and author-provided keywords reduced the dataset to 142 records. Of these, eight articles could not be retrieved in full text despite repeated access attempts through institutional subscriptions and alternative publisher platforms and were therefore excluded at the eligibility stage. leaving 134 studies for full-text eligibility assessment. Following full-text screening, 103 studies were retained for final inclusion, as several articles did not fully meet the predefined inclusion criteria. Using Haddaway et al.'s (2022) PRISMA flowchart tool, Figure 1 below describes the process.

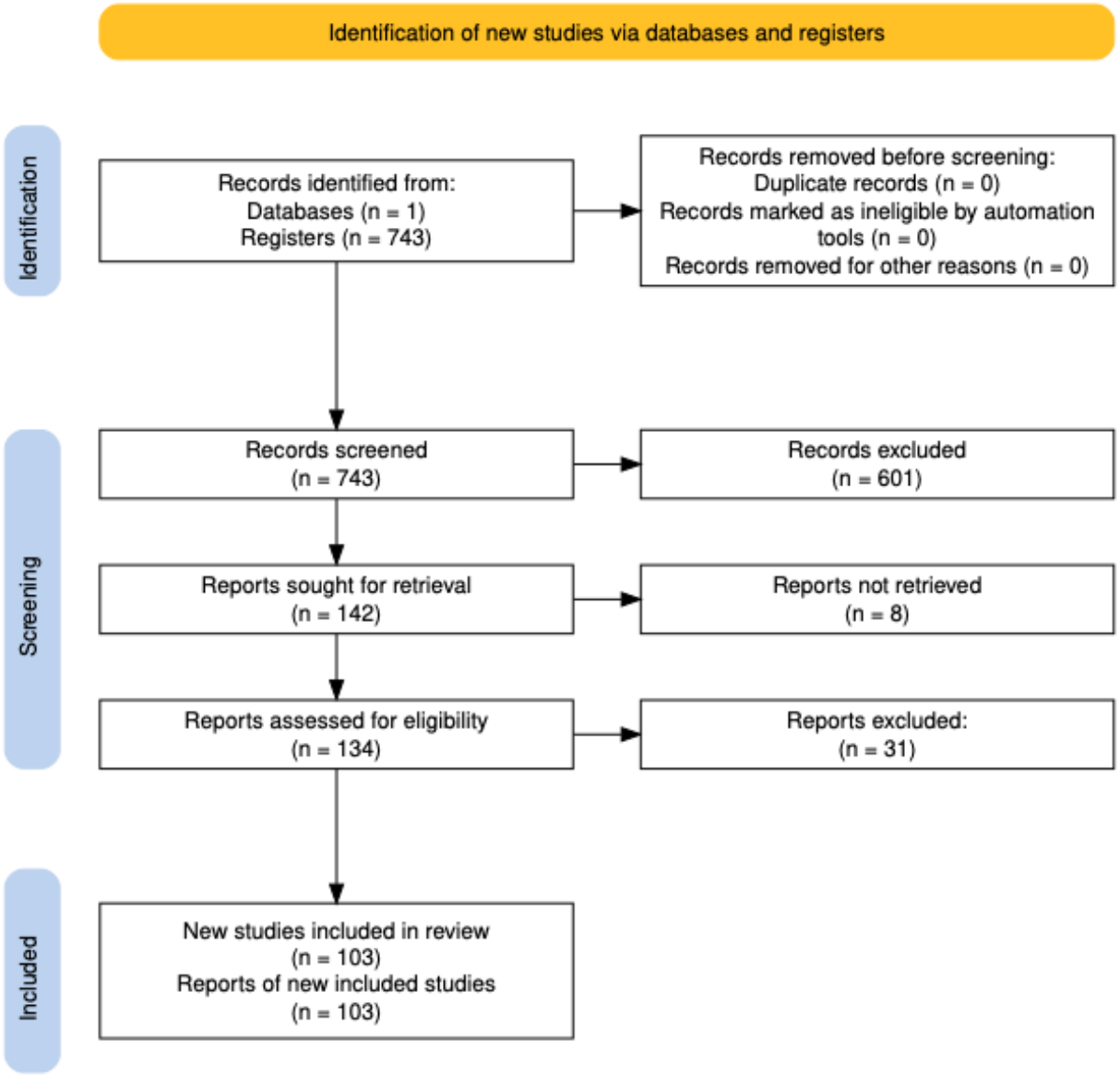

Fig. 1. PRISMA flowchart.

3.5. Quality Assessment

The current study followed the principle that systematic reviews must assess the rigor and relevance of included studies to support trustworthy synthesis, as emphasized by Gough et al. (2017). Studies were evaluated for clarity of design, adequacy of reporting, appropriateness of vocabulary measures, transparency of multimodal manipulation, and alignment between research questions and methods. Rather than excluding studies based solely on quality scores, a weight-of-evidence approach was adopted, whereby methodological limitations informed the interpretation of findings.

3.6. Ethical Considerations

As this review synthesized previously published studies, no institutional ethical approval was required. However, the review adhered to ethical principles of transparency and responsible scholarship, consistent with the expectations outlined in PRISMA 2020 (Page et al., 2021) and the framing of systematic reviews as accountable research processes in Gough et al. (2017).

4. Results

4.1. Theorizing Video-Based Multimodality as a Mechanism for Incidental Vocabulary Acquisition

Across the studies reviewed, theorization converges on a cognitive-processing account in which audiovisual input, augmented by on-screen text and optional glosses, strengthens lexical encoding by distributing information across channels. Dual Coding Theory (Paivio, 2007) and Mayer's Cognitive Theory of Multimedia Learning (Mayer, 2009) dominate as the core explanatory frameworks, with studies treating multimodality as a means of constructing richer form–meaning traces via coordinated verbal (audio/text) and nonverbal (visual/action) representations (Feng & Webb, 2020; Teng & Cui, 2025a, 2025b; Wang, 2025; Yuan & Tang, 2025). Within this framing, captions/subtitles are theorized not merely as redundant text, but as scaffolding that stabilizes transient speech, improves segmentation of the speech stream, and supports mapping between phonological form and semantic representation (Hsu et al., 2013; Wu et al., 2022; Yüksel & Tanrıverdi, 2009).

A second, partially competing theoretical strand centers on load management. Cognitive Load Theory is mobilized to explain both benefits and boundary conditions: multimodality can optimize resources through channel distribution, yet simultaneous presentation of audio, moving images, and dense textual support can trigger split attention and overload, particularly for lower-proficiency learners or high-density captioning conditions (Baranowska, 2025; Choi, 2023; Kaderoğlu, 2024; Montero Perez et al., 2015). This tension appears in the redundancy debate: some studies anticipate a redundancy effect (extra text increases extraneous load), whereas many L2-focused accounts predict a "reversed redundancy effect," arguing that captions reduce intrinsic load by externalizing decoding demands and freeing capacity for meaning construction.

Engagement-oriented models provide a third layer of theorization. Noticing is framed as a mechanism by which multimodal cues, typographic enhancement, and gloss affordances direct attentional resources toward target lexical items, thereby increasing the likelihood of intake and subsequent consolidation (Choi, 2023; Finger-Bou & Muñoz, 2023; Montero Perez et al., 2015; Wu et al., 2022). In parallel, the Involvement Load Hypothesis (Hulstijn & Laufer, 2001) and Depth of Processing Hypothesis (Craik & Lockhart, 1972) are used to account for differences across task types: production-oriented or evaluation-requiring activities are theorized to generate deeper processing and more durable learning than passive viewing, particularly for vocabulary depth and productive outcomes (Lo, 2024; Mohsen, 2016a). Taken together, the accumulated theorization positions incidental vocabulary learning in video-based settings as an emergent product of (a) dual-channel encoding, (b) cognitive load calibration, and (c) attention/engagement mechanisms that vary by the salience and task-demands attached to lexical items.

4.2. Cognitive, Affective, and Behavioral Mediators of Uptake and Retention

4.2.1. Cognitive Mediators

Attention allocation is consistently treated as a necessary gateway to learning in video contexts. Eye-tracking evidence indicates that adding captions/subtitles systematically redistributes gaze from imagery toward text, implying a trade-off that can either facilitate learning (via increased lexical noticing) or undermine processing of visual context if text dominates (Choi, 2023; Montero Perez et al., 2015; Wang, 2025). Beyond "looking," depth of processing is repeatedly linked to activities that require semantic evaluation, contextual interpretation, or retrieval, which promote stronger retention than exposure alone (Çekiç, 2024; Lo, 2024; Mohsen, 2016a; Yuan & Tang, 2025). Working memory capacity appears as a moderator of these effects: higher WM supports coordination of simultaneous audio-text-visual streams and mitigates split-attention costs, especially in complex conditions such as dual subtitles or enhanced captioning (Mirzaei et al., 2023; Teng, 2025). Several accounts also extend mediation beyond single-word learning to phraseological development, proposing that multimodal recurrence can make distributional patterns salient and support statistical learning of multiword units and collocations.

4.2.2. Affective Mediators

Engagement and motivational alignment with content emerges as a stable driver of vocabulary gains. Studies treating genre and interest as consequential report that learners' involvement with documentaries, TV series, or other appealing content increases attention persistence and willingness to sustain exposure, thereby amplifying opportunities for incidental uptake (Teng, 2024). Extramural exposure is framed as a broader affectively-mediated pathway: informal viewing contexts correlate with stronger vocabulary outcomes, plausibly via increased motivation and reduced anxiety.

4.2.3. Behavioral Mediators

Self-regulation through learner control functions (pause/rewind/replay, "help options," glossary checks) is repeatedly linked to improved retention. These behaviors appear to operate through two proximal mechanisms: managing transient cognitive load (slowing down input to a processable rate) and enabling repeated attention to target items, which supports consolidation (Montero Perez et al., 2015; Nguyen & Boers, 2019; Wu et al., 2021). Repetition, whether embedded in extensive viewing or induced through re-viewing/segmenting, functions as a key behavioral condition for long-term retention (Majuddin et al., 2021; Rodgers & Webb, 2017). Finally, production-oriented behaviors (retelling, sentence writing, subtitling/dubbing, content creation) are framed as mediators that convert exposure into elaboration and retrieval practice, thereby strengthening depth of word knowledge (Lo, 2024; Mohsen, 2016a).

4.3. Operationalization and Measurement of Incidental Learning, and How These Choices Condition RQ2 Processes

Across studies, operationalizations vary substantially, and these choices shape which mediating processes become observable and which learning outcomes are likely to emerge. Studies operationalize "incidental" learning through designs ranging from controlled experiments with specific captioning/gloss conditions to more naturalistic viewing contexts that approximate authentic consumption (Lin, 2010; Wang, 2025; Younas & Dong, 2024). Exposure regimes span intensive short-clip designs and extensive viewing designs, with extensive exposure positioned as more plausible for meeting frequency thresholds required for stable retention, particularly beyond initial form recognition.

Measurement timing functions as a strong conditioning factor. Immediate post-tests often show advantages for captioned or bilingual conditions, particularly for recognition and short-term meaning access, whereas delayed tests frequently show attenuation of gains unless repetition is built into the design (Peters, 2019; Teng, 2023; Yüksel & Tanrıverdi, 2009). This pattern aligns with the mediator account in RQ2: immediate designs privilege attention allocation and initial noticing effects, whereas delayed designs are more sensitive to consolidation processes that depend on repetition, self-regulation, and retrieval/production opportunities.

Outcome constructs also condition conclusions. Many studies emphasize receptive breadth (form recognition, meaning recognition/recall), using instruments such as form-meaning tests and VKS-style measures, which tends to favor claims that video/captioning is effective because these outcomes are tightly linked to noticing and surface mapping processes. By contrast, measures targeting depth (productive use, contextualized meaning, collocations/MWUs) more often require enhanced input or task-induced elaboration, consistent with involvement load and depth-of-processing accounts (Lo, 2024; Mohsen, 2016a). In practical terms, the evidence base is partially "measurement-shaped": designs optimized for recognition outcomes will foreground caption-driven attention effects, whereas designs including production, delayed assessment, or phraseological measures will foreground self-regulation, repetition, and elaborative processing as decisive mechanisms.

4.4. Design Principles and Actionable Implications, and the Conditions Under Which They Are Most Plausible

4.4.1. Captioning/subtitling Configurations

The synthesized pattern indicates differentiated affordances rather than a single "best" option. L2 (intralingual) captions are the most consistently supported configuration for linking sound and print, supporting form recognition and listening development via synchronized phonological-orthographic mapping. Bilingual/dual subtitles tend to advantage immediate comprehension and meaning access, particularly for lower-proficiency learners, but risk increased load and reduced form-focused processing if attention is captured by L1 text (Kaderoğlu, 2024; Wang, 2025; Wi & Boers, 2024; Yuan & Tang, 2025). L1-only subtitles reliably support plot comprehension yet appear least aligned with L2 form learning. To mitigate cognitive load in bilingual conditions, sequential strategies, such as initial viewing with L1 subtitles followed by bilingual subtitles, are presented as plausible for preserving comprehension while enabling later lexical focus (Yuan & Tang, 2025).

4.4.2. Salience Engineering via Enhancement and Glossing

Enhanced/keyword captions and typographic highlighting consistently function as attention-guidance tools that intensify noticing without necessarily increasing processing demands to the same extent as full dual-subtitle input (Choi, 2023; Finger-Bou & Muñoz, 2023). Glossing evidence points toward multimodal glosses (e.g., text + picture) as a principled instantiation of dual coding that supports retention beyond single-mode glosses. Interactive glosses (e.g., multiple-choice formats) appear promising, but outputs flag that effectiveness likely depends on feedback timing, visual support, and the cognitive operations invoked during viewing (Çekiç, 2024).

4.4.3. Pacing, Segmenting, and Repetition

Segmenting and pacing are presented as load-management strategies that improve comprehension and prevent overload, particularly for low-proficiency learners or dense captioning environments (Baranowska, 2025). Repetition is treated as non-negotiable for durable learning: evidence summarized here suggests that incidental learning from series-like input may require more than ten encounters for meaning recall to improve reliably, implying that instructional designs should incorporate spaced re-viewing rather than relying on massed exposure alone (Rodgers & Webb, 2017).

4.4.4. Learner Control and Task Sequencing

Providing control tools (pause/rewind/replay) is a consistent implication, framed as enabling self-regulation and adaptive load management that supports deeper processing of target items (Nguyen & Boers, 2019; Wu et al., 2021). Sequencing tasks from lower to higher involvement, for instance passive viewing progressing to evaluation/production activities, is theorized and evidenced as a pathway to vocabulary depth and retention (Lo, 2024; Mohsen, 2016b). The most defensible implication is an integrated model: extensive viewing to accumulate encounters and contextual richness, combined with targeted intentional follow-up on selected items through glossed review, retrieval practice, and production-oriented tasks.

4.4.5. Conditions for Plausibility

Proficiency level and WM capacity emerge as boundary conditions, with intermediate learners and higher-WM learners better positioned to benefit from complex multimodal configurations (Mirzaei et al., 2023; Teng, 2024). Content engagement and genre suitability also condition effectiveness; evidence suggests that motivational fit supports sustained attention and repeated exposure, increasing the plausibility of incidental uptake and retention (Teng, 2024).

5. Discussion

The current study portrayed incidental vocabulary learning under viewing conditions as a multi-component phenomenon rather than an input-format effect. Across studies, the dominant explanatory architecture combined (a) dual-channel representational enrichment, (b) cognitive load calibration, and (c) attention and engagement mechanisms that differed by captioning configuration, salience engineering, and task demands. This triadic account reduced apparent contradictions in the evidence base: captions and multimodal enhancements supported form-meaning encoding, yet gains varied because the same supports also reallocated attention and could exceed processing capacity in high-density or complex subtitle conditions. In practical terms, multimodality functioned as an enabling condition whose benefits depended on the extent to which designs stabilized transient speech, guided selective attention to target items, and preserved manageable load.

Within the findings, the strongest internal convergence concerned mediation. Attention allocation emerged as a necessary gateway, with eye-tracking studies indicating systematic shifts of gaze toward captions/subtitles, implying that more text support rarely represented a monotonic improvement. Depth of processing and self-regulation then differentiated short-term uptake from durable retention. Learner control behaviors, pause/rewind/replay, and repeated encounters consistently aligned with longer-term outcomes, while production-oriented activities plausibly converted exposure into elaboration and retrieval practice. This pattern also explained the repeated immediate acquisitions over delayed ones reported across studies. Immediate measures privileged noticing and initial mapping, while delayed measures required consolidation conditions, repetition, and post-viewing operations that were absent in many protocols.

A second internal convergence concerned measurement-shaped conclusions. The evidence base leaned heavily toward receptive breadth outcomes (form and meaning recognition/recall), which structurally favored claims about caption-driven noticing and surface mapping. Designs incorporating productive measures, phraseological outcomes, or delayed assessment foregrounded involvement load, repetition, and elaborative processing as decisive mechanisms. This asymmetry suggested that the field's effects reflected not only learning processes but also the psychometric and temporal windows selected for observation.

These patterns aligned tightly with the theoretical framing. Dual Coding Theory and the Cognitive Theory of Multimedia Learning provided the backbone for interpreting multimodal enrichment, while the noticing account clarified why typographic enhancement and captioning increased the probability of intake through attention guidance (Montero Perez et al., 2014, 2018; Sydorenko, 2010; Warren et al., 2018). The findings extended that framing by making the trade-off explicit. Captioning increased lexical attention, yet excessive text or dual subtitles plausibly displaced attention from visual context, thereby altering the informational basis for inference and semantic integration. The cognitive load strand in your review also cohered with the results' boundary conditions: multimodality supported learning under calibrated designs but risked split attention and overload under dense or complex configurations, particularly for lower-proficiency learners (Hsu et al., 2013; Ko, 2012, 2017; Nation, 2013). Finally, the findings' emphasis on task sequencing and production-oriented follow-up resonated with the involvement load and depth-of-processing logic already present in the literature review (Vahedi et al., 2016; Yun, 2011), while specifying that these mechanisms became most visible under depth-sensitive outcomes.

Comparison with the two earlier reviews further sharpened the contribution of the present synthesis. Simonnet et al. (2025) characterized technology-assisted vocabulary learning as generally beneficial for vocabulary outcomes and learner experience, while highlighting methodological biases and the difficulty of generalizing across heterogeneous protocols; they explicitly recommended more standardized experimental procedures and comparable measures to support generalization. The current findings echoed this concern through the measurement-shaped finding and the repeated sensitivity of conclusions to

testing timing, outcome construct, and exposure regime. Simonnet et al. (2025) also emphasized novelty and attention-to-vocabulary as plausible confounds and noted the prevalence of quasi-experimental designs and ad hoc instruments as barriers to causal inference and comparability, which mapped directly onto the present review's call to interpret captioning/gloss effects through the lens of design choices that instantiate attention, repetition, and self-regulation.

Zeng et al. (2025) approached vocabulary instruction through an explicit theory-to-practice lens and foregrounded schema/psycholinguistic, sociocultural, motivational, and dual coding frameworks as common guides for instructional work. This theoretical palette overlapped with the present synthesis. Dual coding and multimedia learning explained representational enrichment, while motivation and engagement models cohered with the results showing content interest and extramural viewing as stable amplifiers of sustained attention and exposure. In addition, Zeng et al. (2025) stressed multiple exposures, interactive learning opportunities, and the use of visuals and high-interest multimedia texts as supports for vocabulary development, which aligned with the present design principles emphasizing repetition, learner control tools, and integrated sequencing from extensive viewing to targeted, higher-involvement follow-up.

6. Conclusion

This systematic review synthesized evidence on English vocabulary acquisition under viewing conditions and clarified how video-based multimodality plausibly supported incidental learning. The reviewed studies converged on an account in which learning emerged from the interaction of three mechanism families rather than from modality alone: (a) dual-channel representational enrichment (audiovisual input combined with on-screen text and optional glosses), (b) cognitive load calibration (benefits under manageable density and segmentation, costs under split attention and overload), and (c) attention and engagement processes that varied by captioning configuration, salience engineering, and task demands. Within that integrated account, captions and related enhancements functioned as stabilizers of transient speech and as attention-guidance cues that increased opportunities for form–meaning mapping, yet gains depended on whether the design preserved processing capacity and enabled repeated, self-regulated engagement.

The synthesis also demonstrated that conclusions in this area have been "measurement-shaped." Many studies emphasized receptive breadth outcomes (e.g., recognition-based measures) and immediate post-tests, which tended to foreground noticing and initial mapping effects. Designs incorporating delayed assessment, productive outcomes, or phraseological measures more often highlighted repetition, learner control (pause/rewind/replay and help options), and production-oriented follow-up as decisive for consolidation and depth of word knowledge. Proficiency and working memory capacity operated as boundary conditions, with more complex multimodal configurations (e.g., dual subtitles and dense textual support) appearing more plausible for intermediate or higher-capacity learners, while engagement with content and genre suitability amplified exposure and persistence.

7. Limitations and Suggestions for Further Research

At the instructional-design level, the evidence supports differentiated, purpose-driven use of captioning and salience tools rather than a single "best" configuration. Intralingual (L2) captions aligned with phonological-orthographic mapping and lexical noticing, while bilingual subtitles often supported immediate comprehension and meaning access but risked load inflation and reduced L2 form focus if attention was captured by L1 text. This pattern supports sequencing logic, for example, comprehension-first viewing followed by more form-focused configurations and tasks, to balance meaning construction with lexical encoding. Salience engineering through keyword captions, typographic enhancement, and multimodal glossing remained defensible primarily as attention-guidance devices that should be tuned to avoid density-based overload. Crucially, repetition and learner control features emerged as non-negotiable

for durable learning, indicating that viewing-based vocabulary work benefits from designs that incorporate re-viewing, segmenting, and spaced encounters rather than single-pass exposure.

Methodologically, the present synthesis reinforced wider review-level concerns about heterogeneity and comparability. Simonnet et al. (2025) emphasized that varied measurement methods, frequent quasi-experimental designs without control groups, and heavy reliance on ad hoc instruments constrained causal inference and generalization, while standardized tests and more comparable protocols would improve comparability across experiments. Those cautions directly align with the present review's identification of timing- and construct-dependence, and they strengthen the implication that future evidence should be interpreted through explicit links between mechanism, operationalization, and outcome construct.

From a theory-to-practice perspective, the findings fit within broader instructional frameworks emphasized in practitioner-oriented syntheses. Zeng et al. (2025) highlighted that vocabulary instruction has been commonly anchored in schema/psycholinguistic theories, sociocultural perspectives, motivation, and dual coding, and they anticipated increased use of multimodal approaches such as video and gamification in contemporary practice. The present review refines that expectation by specifying mechanism-contingent conditions under which multimodal viewing is most plausible for incidental vocabulary growth: calibrated load, guided attention, repeated encounters, and structured opportunities for elaboration and retrieval.

Several limitations should be acknowledged when interpreting these findings. First, the review was intentionally restricted to the Web of Science Core Collection to ensure metadata consistency and analytical coherence for ecological mapping; however, this decision may have excluded relevant studies published in journals indexed exclusively in other databases. As such, the synthesis prioritizes conceptual density and pattern stability over maximal recall. Second, a small number of studies (n = 8) could not be retrieved in full text despite repeated access attempts and were excluded at the eligibility stage. Although this represents a minor proportion of the initial corpus, their exclusion introduces a limited retrieval bias that cannot be entirely ruled out. Third, as noted above, substantial heterogeneity in design quality, outcome operationalization, and assessment timing constrains cross-study comparability and limits causal inference. Consequently, the conclusions should be interpreted as mechanism-oriented tendencies rather than definitive effect claims.

Future studies should prioritize (1) standardized and comparable measurement practices, including designs that allow stronger causal attribution and cross-study comparison, consistent with recommendations for standardized protocols and measures in the technology-assisted vocabulary literature. (2) Mechanism-sensitive designs that align outcomes with hypothesized processes, for example, delayed testing and productive or phraseological measures when consolidation and depth are central claims. (3) Transparent operationalization of multimodal conditions, including subtitle density, enhancement parameters, segmentation and pacing, and the affordances for learner control, to clarify boundary conditions rather than treating multimodality as a unitary treatment. (4) Multi-method mediation approaches, combining outcome tests with attention and interaction indicators (e.g., eye-tracking and trace analyses), which Simonnet et al. (2025) framed as valuable for observing behavior over time and strengthening interpretation beyond pre/post contrasts. Collectively, these steps should move the field from "effects of captions/videos" toward falsifiable, mechanism-aligned claims about how, for whom, and under what viewing configurations incidental vocabulary learning becomes durable and instructionally usable.

References

Akbulut, Y. (2007). Effects of multimedia annotations on incidental vocabulary learning and reading comprehension of advanced learners of English as a foreign language. *Instructional Science, 35*(6), 499–517. https://doi.org/10.1007/s11251-007-9016-7

Aldera, A. S., & Mohsen, M. A. (2013). Annotations in captioned animation: Effects on vocabulary learning and listening skills. *Computers & Education, 68*, 60–75. http://dx.doi.org/10.1016/j.compedu.2013.04.018

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

Baranowska, K. (2025). What watching subtitled movies does to the learner: The impact of subtitles and modality on cognitive load and vocabulary learning. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2024-0175

Boers, F., Warren, P., Grimshaw, G., & Siyanova-Chanturia, A. (2017b). On the benefits of multimodal annotations for vocabulary uptake from reading. *Computer Assisted Language Learning, 30*, 709–725. https://doi.org/10.1080/09588221.2017.1356335

Boers, F., Warren, P., He, L., & Deconinck, J. (2017a). Does adding pictures to glosses enhance vocabulary uptake from reading? *System, 66*, 113–129. https://doi.org/10.1016/j.system.2017.03.017

Choi, S. (2023). Visual saliency in captioned digital videos and learning of English collocations: An eye-tracking study. *Language Learning & Technology, 27*(1), 1–21. https://doi.org/10.64152/10125/73536

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

Çekiç, A. (2024). Incidental L2 vocabulary learning from audiovisual input: The effects of different types of glosses. *Computer Assisted Language Learning, 37*(4), 896–923. https://doi.org/10.1080/09588221.2022.2062004

Fakhr, M. A., Borzabadi Farahani, D., & Khomeijani Farahani, A. A. (2021). Incidental vocabulary learning and retention from audiovisual input and factors affecting them. *English Teaching & Learning, 45*(2), 167–188. https://doi.org/10.1007/s42321-020-00066-y

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition, 42*(3), 499–523. https://doi.org/10.1017/S0272263119000494

Fievez, I., Montero Perez, M., Cornillie, F., & Desmet, P. (2023). Promoting incidental vocabulary learning through watching a French Netflix series with glossed captions. *Computer Assisted Language Learning, 36*(1-2), 26–51. https://doi.org/10.1080/09588221.2021.1899244

Finger-Bou, R., & Muñoz, C. (2023). The effects of regular and enhanced captions on incidental vocabulary acquisition. *ELIA: Estudios de Lingüística Inglesa Aplicada, 23*, 15–50. https://doi.org/10.12795/elia.2023.i23.01

Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words. *Studies in Second Language Acquisition, 35*(3), 483–517. https://doi.org/10.1017/s0272263113000119

Gough, D., Oliver, S., & Thomas, J. (2017). *An introduction to systematic reviews*. SAGE.

Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and open synthesis. *Campbell Systematic Reviews, 18*(2). https://doi.org/10.1002/cl2.1230

Hsieh, Y. (2020). Effects of video captioning on EFL vocabulary learning and listening comprehension. *Computer Assisted Language Learning, 33*(5–6), 567–589. https://doi.org/10.1080/09588221.2019.1577898

Hsu, C.-K., Hwang, G.-J., Chang, Y.-T., & Chang, C.-K. (2013). Effects of video caption modes on English listening comprehension and vocabulary acquisition using handheld devices. *Educational Technology & Society, 16*(1), 403–414.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*(3), 539–558. https://doi.org/10.1111/0023-8333.00164

Kaderoğlu, K. (2024). Incidental vocabulary learning from audiovisual input: The case of pre-intermediate Turkish EFL learners. *ELIA*, (24), 177–207. https://doi.org/10.12795/elia.2024.i24.6

Ko, M. H. (2012). Glossing and second language vocabulary learning. *TESOL Quarterly, 46*, 56–79. https://doi.org/10.1002/tesq.3

Ko, M. H. (2017). The relationship between gloss type and L2 proficiency in incidental vocabulary learning. *The Modern English Society, 18*, 47–69. https://doi.org/10.18095/meeso.2017.18.3.03

Lee, T., & Choi, S. (2024). Glossed video keyword captions and L2 vocabulary acquisition: An eye-tracking study. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2024.2412103

Lin, L.-F. (2010). A video-based CALL program for proficient and less-proficient L2 learners' comprehension ability, incidental vocabulary acquisition. *Educational Media International, 47*(3), 199–216. https://doi.org/10.1080/09523987.2010.518812

Lo, S. (2024). Learning vocabulary through dual-subtitled viewing: The impact of different ILH-based interventions. *Computer Assisted Language Learning, 37*(7), 1829–1856. https://doi.org/10.1080/09588221.2022.2126497

Majuddin, E., Siyanova-Chanturia, A., & Boers, F. (2021). Incidental acquisition of multiword expressions through audiovisual materials: The role of repetition and typographic enhancement. *Studies in Second Language Acquisition, 43*(4), 985–1008. https://doi.org/10.1017/S0272263121000036

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

Mayer, R. E. (2009). *Multimedia learning* (2nd ed.). Cambridge University Press.

Mirzaei, A., Azizi Farsani, M., & Chang, H. (2023). Statistical learning of L2 lexical bundles through unimodal, bimodal, and multimodal stimuli. *Language Teaching Research*, 1–25. https://doi.org/10.1177/13621688231193079

Mohsen, M. A. (2016a). The use of computer-based simulation to aid comprehension and incidental vocabulary learning. *Journal of Educational Computing Research, 54*(6), 863–884. https://doi.org/10.1177/0735633116639954

Mohsen, M. A. (2016b). Effects of help options in a multimedia listening environment on L2 vocabulary acquisition. *Computer Assisted Language Learning, 29*(7), 1220–1240. https://doi.org/10.1080/09588221.2015.1093175

Montero Perez, M., Peters, E., & Desmet, P. (2015). Enhancing vocabulary learning through captioned video: An eye-tracking study. *The Modern Language Journal, 99*(2), 308–328. https://doi.org/10.1111/modl.12215

Montero Perez, M., Peters, E., & Desmet, P. (2018). Vocabulary learning through viewing video: The effect of two enhancement techniques. *Computer Assisted Language Learning*, *31*(1–2), 1–26. https://doi.org/10.1080/09588221.2017.1375960

Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary. *Language, Learning & Technology, 18*(1), 118–141. http://dx.doi.org/10.10125/44357

Nagata, N. (1999). The effectiveness of computer-assisted interactive glosses. *Foreign Language Annals, 32*, 469–479. https://doi.org/10.1111/j.1944-9720.1999.tb00876.x

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press

Nguyen, C.-D., & Boers, F. (2019). The effect of content retelling on vocabulary uptake from a TED Talk. *TESOL Quarterly, 53*(1), 5–29. https://doi.org/10.1002/tesq.441

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hrobjartsson, ´ A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Paivio, A. (2007). *Mind and its evolution: A dual coding theoretical approach*. Lawrence Erlbaum Associates Publishers.

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly, 53*(4), 1008–1032. https://doi.org/10.1002/tesq.531

Peters, E., Heynen, E., & Puimège, E. (2016). Learning vocabulary through audiovisual input: The differential effect of L1 subtitles and captions. *System, 63*, 134–148. https://doi.org/10.1016/j.system.2016.10.002

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.

Rassaei, E. (2018). Computer-mediated textual and audio glosses, perceptual style and L2 vocabulary learning. *Language Teaching Research, 22*(6), 657–675. https://doi.org/10.1177/1362168817690183

Rodgers, M. P. H., & Webb, S. (2020). IVL through viewing television. *ITL - International Journal of Applied Linguistics, 171*(2), 191–220. https://doi.org/10.1075/itl.18034.rod

Rodgers, M. P. H., & Webb, S. (2017). The effects of captions on EFL learners' comprehension of English-language television programs. *CALICO Journal, 34*(1), 20–38. https://doi.org/10.1558/cj.29522

Salehi, V., & Naserieh, F. (2013). The effects of verbal glosses on vocabulary learning and reading comprehension. *Asian EFL Journal, 15*, 24–64.

Simonnet, E., Loiseau, M., & Lavoué, É. (2025). A systematic literature review of technology-assisted vocabulary learning. *Journal of Computer Assisted Learning, 41*(1). https://doi.org/10.1111/jcal.13096

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

Teng, M. F. (2020). Retention of new words learned incidentally from reading: Word exposure frequency, L1 marginal glosses, and their combination. *Language Teaching Research, 24*(6), 785–812. https://doi.org/10.1177/1362168819829026

Teng, M. F. (2023). The effectiveness of multimedia input on vocabulary learning and retention. *Innovation in Language Learning and Teaching, 17*(3), 738–754. https://doi.org/10.1080/17501229.2022.2131791

Teng, M. F. (2024). Incidental vocabulary learning from captioned video genres: Proficiency, working memory, and aptitude. *Computer Assisted Language Learning*, 1–43. https://doi.org/10.1080/09588221.2024.2421517

Teng, M. F. (2025). Incidental vocabulary learning from captioned video genres: Vocabulary knowledge, comprehension, repetition, and working memory. *Computer Assisted Language Learning, 38*(5–6), 1301–1340. https://doi.org/10.1080/09588221.2023.2275158

Teng, M. F., & Cui, Y. (2025a). Incidental vocabulary learning from captioned viewing: Relative contributions of word- and learner-related factors. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688251372987

Teng, M. F., & Cui, Y. (2025b). Second language collocation learning through captioned videos: How do learners' vocabulary knowledge and working memory affect learning? *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2025.2497495

Vahedi, V. S., Ghonsooly, B., & Pishghadam, R. (2016). Vocabulary glossing: A meta-analysis of the relative effectiveness of different gloss types on L2 vocabulary acquisition. *Teaching English with Technology, 16*, 3–25.

Wang, A. (2025). The integration of auditory and textual input in vocabulary learning from subtitled viewing: An eye-tracking study. *Language Learning & Technology, 29*(3), 70–91. https://doi.org/10.64152/10125/73648

Warren, P., Boers, F., Grimshaw, G., & Siyanova-Chanturia, A. (2018). The effect of gloss type on learners' intake of new words during reading: Evidence from eye-tracking. *Studies in Second Language Acquisition, 40*, 883–906. https://doi.org/10.1017/S0272263118000177

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Webb, S., & Rodgers, M. P. H. (2009). The lexical coverage of movies. *Applied Linguistics, 30*, 407 – 427. http://doi.org/10.1093/applin/amp010

Wi, I., & Boers, F. (2024). Sequential use of L1 and L2 captions: Exploring the benefits for vocabulary acquisition. *TESOL Quarterly, 58*(1), 511–531. https://doi.org/10.1002/tesq.3243

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology, 14*(1), 65–86.

Wu, H., Yu, P., Yang, S., & Chen, X. (2022). Video captioning effects on EFL listening comprehension and vocabulary learning. *International Journal of Computer-Assisted Language Learning and Teaching, 12*(2), 1–16. https://doi.org/10.4018/ijcallt.291534

Wu, W.-C. V., Lin, I.-T. D., Marek, M. W., & Ou Yang, F.-C. (2021). Analysis of English idiomatic learning behaviors of an audio-visual mobile application. *SAGE Open, 11*(2), 1–17. https://doi.org/10.1177/21582440211016899

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

Yeldham, M. (2018). Viewing L2 captioned videos: What's in it for the listener? *Computer Assisted Language Learning, 31*(4), 367–389. https://doi.org/10.1080/09588221.2017.1406956

Younas, M., & Dong, Y. (2024). The impact of using animated movies in learning English language vocabulary: An empirical study of Lahore, Pakistan. *SAGE Open, 14*(2), 1–12. https://doi.org/10.1177/21582440241258398

Yuan, X., & Tang, X. (2025). Effects of the sequential use of L1 and bilingual subtitles on incidental English vocabulary learning: A cognitive load perspective. *British Journal of Educational Psychology, 95*(2), 565–577. https://doi.org/10.1111/bjep.12740

Yüksel, D., & Tanrıverdi, B. (2009). Effects of watching captioned movie clips on vocabulary development of EFL learners. *The Turkish Online Journal of Educational Technology, 8*(2).

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning, 24*, 39–58. https://doi.org/10.1080/09588221.2010.523285

Zeng, Y., Kuo, L.-J., Chen, L., Lin, J.-A., & Shen, H. (2025). Vocabulary instruction for English learners: A systematic review connecting theories, research, and practices. *Education Sciences, 15*(3), 262. https://doi.org/10.3390/educsci15030262

Appendix A. Studies Included in the Systematic Review

Aldera, A. S., & Mohsen, M. A. (2013). Annotations in captioned animation: Effects on vocabulary learning and listening skills. *Computers & Education, 68*, 60–75. http://dx.doi.org/10.1016/j.compedu.2013.04.018

Alshumrani, H. (2023). The learning potential of a TV series in promoting L2 incidental learning of idiomatic and non-idiomatic phrasal verbs. *Journal of Language and Education, 9*(3), 12–23. https://doi.org/10.17323/jle.2023.17302

Ansarin, A. A., & Khabbazi, S. K. (2021). Task-induced involvement load and working memory: Effects on active and passive vocabulary knowledge of EFL learners in a multimedia learning environment. *Eurasian Journal of Applied Linguistics, 7*(1), 277–302. https://doi.org/10.32601/ejal.911288

Aziz, Z. A., Mustafa, F., & Yulia, M. (2024). English vocabulary retention on movie series with L1 and English subtitles: The role of vocabulary level and frequency. *LLT Journal: A Journal on Language and Language Teaching, 27*(2), 582–598.

Balenovic, K., & Prorokovic, J. (2023). Foreign Language Vocabulary Development: Media–driven learning in the informal context. *Suvremena Lingvistika, 49*(96), 175–201. https://doi.org/10.22210/suvlin.2023.096.01

Baranowska, K. (2020). Learning most with least effort: Subtitles and cognitive load. *ELT Journal, 74*(2), 105–115. https://doi.org/10.1093/elt/ccz060

Baranowska, K. (2025). What watching subtitled movies does to the learner: The impact of subtitles and modality on cognitive load and vocabulary learning. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2024-0175

Bellalem, F., Neddar, B. A., Bouagada, H., & Djelloul, D. B. (2018). The use of subtitled movies for vocabulary acquisition in ESP settings: Insights from an experimental study in Algeria. *Arab World English Journal, 9*(3), 3–16. https://dx.doi.org/10.24093/awej/vol9no3.1

Bobkina, J., Baluyan, S., & Domínguez Romero, E. (2025). Tech-enhanced vocabulary acquisition: Exploring the use of student-created video learning materials in the tertiary-level EFL (English as a Foreign Language) flipped classroom. *Education Sciences, 15*(4), 450. https://doi.org/10.3390/educsci15040450

Cárdenas-Claros, M. S., & Ramírez-Orellana, D. (2024). Progressive reduction of captions in language learning. *Journal of Information Technology Education: Innovations in Practice, 23*, Article 2. https://doi.org/10.28945/5263

Çekiç, A. (2024). Incidental L2 vocabulary learning from audiovisual input: The effects of different types of glosses. *Computer Assisted Language Learning, 37*(4), 896–923. https://doi.org/10.1080/09588221.2022.2062004

Çekiç, A., & Demirezen, M. (2020). Comparison of the impacts of different multimodalities on incidental L2 vocabulary learning. *Moderna Språk, 114*(2), 109–138. https://doi.org/10.58221/mosp.v114i2.7405

Chen, C.-M., Li, M.-C., & Lin, M.-F. (2020). The effects of video-annotated learning and reviewing system with vocabulary learning mechanism on English listening comprehension and technology acceptance. *Computer Assisted Language Learning, 35*(7), 1557–1593. https://doi.org/10.1080/09588221.2020.1825093

Chen, I.-S. J. (2020). Music as a mnemonic device for foreign vocabulary learning. *English Teaching & Learning, 44*(4), 377–395. https://doi.org/10.1007/s42321-020-00049-z

Chen, S. (2024). Effects of subtitles on Vocabulary learning through videos: An exploration across different learner types. *The Journal of Specialised Translation, 42*(42), 257–276. https://doi.org/10.26034/cm.jostrans.2024.5992

Chen, S. (2025). Subtitles for vocabulary learning: Assessing the effects of L2, L1, and bilingual subtitles over time. *System, 132*, 103709. https://doi.org/10.1016/j.system.2025.103709

Chen, Y.-R., Liu, Y.-T., & Todd, A. G. (2018). Transient but effective? Captioning and adolescent EFL learners' spoken vocabulary acquisition. *English Teaching & Learning, 42*(1), 25–56. https://doi.org/10.1007/s42321-018-0002-8

Choi, S. (2023). Visual saliency in captioned digital videos and learning of English collocations: An eye-tracking study. *Language Learning & Technology, 27*(1), 1–21. https://doi.org/10.64152/10125/73536

Danan, M. (1992). Reversed subtitling and dual coding theory: New directions for foreign language instruction. *Language Learning, 42*(4), 497–527.

Dang, T. N. Y., Lu, C., & Webb, S. (2022a). Incidental learning of collocations in an academic lecture through different input modes. *Language Learning, 72*(3), 728–764. https://doi.org/10.1111/lang.12499

Dang, T. N. Y., Lu, C., & Webb, S. (2022b). Incidental learning of single words and collocations through viewing an academic lecture. *Studies in Second Language Acquisition, 44*(3), 708–736. https://doi.org/10.1017/S0272263121000474

Dang, T. N. Y., Lu, C., & Webb, S. (2023). Open access academic lectures as sources for incidental vocabulary learning: Examining the role of input mode, frequency, type of vocabulary, and elaboration. *Applied Linguistics, 44*(4), 747–770. https://doi.org/10.1093/applin/amac044

De Wilde, V. (2025). Adolescent learners' L2 English vocabulary knowledge and contact with extramural English: Longitudinal Development and relationships between L2 vocabulary and extramural English. *Bilingualism: Language and Cognition*, 1–12. https://doi.org/10.1017/s1366728925100540

Dolgunsöz, E. (2019). The effect of taboo content on incidental vocabulary acquisition in a foreign language: A facial expression analysis study. *Studia Psychologica, 61*(1), 3–16. https://doi.org/10.21909/sp.2019.01.768

Ebadi, S., Amini, Z., & Gheisari, N. (2023). On the relationship between mobile-based extramural activities and vocabulary development of EFL learners: A mixed-method study. *Smart Learning Environments, 10*(1). https://doi.org/10.1186/s40561-023-00252-y

Etemadi, A. (2012). Effects of bimodal subtitling of English movies on content comprehension and vocabulary recognition. *International Journal of English Linguistics, 2*(1). https://doi.org/10.5539/ijel.v2n1p239

Fainman, I., & Tokar, Y. (2021). Authentic videos in the ESP classroom: Enhancing aviation english vocabulary acquisition. *Argentinian Journal of Applied Linguistics, 9*(2), 52–62.

Fakhr, M. A., Borzabadi Farahani, D., & Khomeijani Farahani, A. A. (2021). Incidental vocabulary learning and retention from audiovisual input and factors affecting them. *English Teaching & Learning, 45*(2), 167–188. https://doi.org/10.1007/s42321-020-00066-y

Fang, F., Zhang, Y., & Fang, Y. (2019). A comparative study of the effect of bilingual subtitles and English subtitles on college english teaching. *Revista de Cercetare Si Interventie Sociala, 66*, 59–74. https://doi.org/10.33788/rcis.66.4

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing: Which mode of input is most effective? *Studies in Second Language Acquisition, 42*(3), 499–523. https://doi.org/10.1017/S0272263119000494

Finger-Bou, R., & Muñoz, C. (2023). The effects of regular and enhanced captions on incidental vocabulary acquisition. *ELIA: Estudios de Lingüística Inglesa Aplicada, 23*, 15–50. https://doi.org/10.12795/elia.2023.i23.01

Frumuselu, A. D., De Maeyer, S., Donche, V., & Colon Plana, M. D. M. G. (2015). Television series inside the EFL classroom: Bridging the gap between teaching and learning informal language through subtitles. *Linguistics and Education, 32*, 107–117. http://dx.doi.org/10.1016/j.linged.2015.10.001

Gesa, F., & Miralpeix, I. (2023). Extensive viewing as additional input for foreign language vocabulary learning: A longitudinal study in secondary school. *Language Teaching Research*. https://doi.org/10.1177/13621688231169451

Haider, A. S., & Al-Salman, S. (2022). The effects of intralingual subtitles on Jordanian university students' foreign language learning. *International Journal of Instruction, 15*(4), 57-76. https://doi.org/10.29333/iji.2022.1544a

Hao, T., Sheng, H., Ardasheva, Y., & Wang, Z. (2021). Effects of dual subtitles on Chinese students' English listening comprehension and vocabulary learning. *The Asia-Pacific Education Researcher, 31*(5), 529–540. https://doi.org/10.1007/s40299-021-00601-w

Hsieh, Y. (2020). Effects of video captioning on EFL vocabulary learning and listening comprehension. *Computer Assisted Language Learning, 33*(5–6), 567–589. https://doi.org/10.1080/09588221.2019.1577898

Hsu, C.-K., Hwang, G.-J., Chang, Y.-T., & Chang, C.-K. (2013). Effects of video caption modes on English listening comprehension and vocabulary acquisition using handheld devices. *Educational Technology & Society, 16*(1), 403–414.

Hsu, W. (2013). The effects of audiovisual support on EFL learners' productive vocabulary. *ReCALL, 26*(1), 62–79. https://doi.org/10.1017/s0958344013000220

Kaderoğlu, K. (2024). Incidental vocabulary learning from audiovisual input: The case of pre-intermediate turkish EFL learners. *ELIA,* (24), 177–207. https://doi.org/10.12795/elia.2024.i24.6

Koolstra, C. M., & Beentjes, J. W. J. (1999). Children's vocabulary acquisition in a foreign language through watching subtitled television programs at home. *Educational Technology Research and Development, 47*(1), 51–60.

Kuppens, A. H. (2010). Incidental foreign language acquisition from media exposure. *Learning, Media and Technology, 35*(1), 65–85. https://doi.org/10.1080/17439880903561876345678 9

Lai, H., Wang, D., & Ou, X. (2021). The effects of different caption modes on Chinese English learners' content and vocabulary comprehension. *International Journal of Computer-Assisted Language Learning and Teaching, 11*(4), 54–68. https://doi.org/10.4018/ijcallt.2021100104

Lee, T., & Choi, S. (2024). Glossed video keyword captions and L2 vocabulary acquisition: An eye-tracking study. *Computer Assisted Language Learning*. https://doi.org/10.1080/09588221.2024.2412103

Li, M., & Hennebry-Leung, M. (2022). Effects of monolingual and bilingual subtitles on L2 vocabulary acquisition. *International Review of Applied Linguistics in Language Teaching, 62*(2), 843–870. https://doi.org/10.1515/iral-2022-0034

Li, W., Yu, J., Zhang, Z., & Liu, X. (2022). Dual coding or cognitive load? exploring the effect of multimodal input on English as a foreign language learners' vocabulary learning. *1*. https://doi.org/10.3389/fpsyg.2022.834706

Lin, L.-F. (2010). A video-based CALL program for proficient and less-proficient L2 learners' comprehension ability, incidental vocabulary acquisition. *Educational Media International, 47*(3), 199–216. https://doi.org/10.1080/09523987.2010.518812

Lo, S. (2024a). Learning vocabulary through dual-subtitled viewing: The impact of different ILH-based interventions. *Computer Assisted Language Learning, 37*(7), 1829–1856. https://doi.org/10.1080/09588221.2022.2126497

Lo, S. (2024b). Vocabulary learning through viewing dual-subtitled videos: Immediate repetition versus spaced repetition as an enhancement strategy. *ReCALL, 36*(2), 152–167. https://doi.org/10.1017/S0958344024000053

Lo, S. (2025). Viewing dual-subtitled videos under different learning conditions: Effects on learners' behavioural, emotional, and cognitive engagement. *Computer Assisted Language Learning*, *38*(4), 742–772. https://doi.org/10.1080/09588221.2023.2219711

Lwo, L., & Lin, M. C.-T. (2012). The effects of captions in teenagers' multimedia L2 learning. *ReCALL*, *24*(2), 188–208. https://doi.org/10.1017/S0958344012000067

Majuddin, E., Boers, F., & Siyanova-Chanturia, A. (2024). The effects of enhancing L2 multiword items in captions: An approximate replication of Majuddin, Siyanova-Chanturia, and Boers (2021). *Language Teaching*, 1–18. https://doi.org/10.1017/S0261444824000296

Majuddin, E., Siyanova-Chanturia, A., & Boers, F. (2021). Incidental acquisition of multiword expressions through audiovisual materials: The role of repetition and typographic enhancement. *Studies in Second Language Acquisition*, *43*(4), 985–1008. https://doi.org/10.1017/S0272263121000036

Mirzaei, A., Azizi Farsani, M., & Chang, H. (2023). Statistical learning of L2 lexical bundles through unimodal, bimodal, and multimodal stimuli. *Language Teaching Research*, 1–25. https://doi.org/10.1177/13621688231193079

Mirzaei, A., Beigi, Z., Eslami, Z. R., Roohani, A., & Burke, M. D. (2025). Effects of L2 multimodality on vocabulary resonance: Full/partial captioning and working memory. *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2025.2515174

Mohsen, M. A. (2016a). The use of computer-based simulation to aid comprehension and incidental vocabulary learning. *Journal of Educational Computing Research*, *54*(6), 863–884. https://doi.org/10.1177/0735633116639954

Mohsen, M. A. (2016b). Effects of help options in a multimedia listening environment on L2 vocabulary acquisition. *Computer Assisted Language Learning*, *29*(7), 1220–1240. https://doi.org/10.1080/09588221.2015.1093175

Mohsen, M. A., & Mahdi, H. S. (2021). Partial versus full captioning mode to improve L2 vocabulary acquisition in a mobile-assisted language learning setting: Words pronunciation domain. *Journal of Computing in Higher Education*, *33*, 524–543. https://doi.org/10.1007/s12528-021-09276-0

Montero Perez, M., Peters, E., & Desmet, P. (2015). Enhancing vocabulary learning through captioned video: An eye-tracking study. *The Modern Language Journal*, *99*(2), 308–328. https://doi.org/10.1111/modl.12215

Muñoz, C., Pujadas, G., & Pattemore, A. (2023). Audio-visual input for learning L2 vocabulary and grammatical constructions. *Second Language Research*, *39*(1), 13–37. https://doi.org/10.1177/02676583211015797

Neuman, S. B., & Koskinen, P. (1992). Captioned television as comprehensible input: Effects of incidental word learning from context for language minority students. *Reading Research Quarterly*, *27*(1), 94–106. https://doi.org/10.2307/747835

Nguyen, C.-D., & Boers, F. (2019). The effect of content retelling on vocabulary uptake from a TED Talk. *TESOL Quarterly, 53*(1), 5–29. https://doi.org/10.1002/tesq.441

Pattemore, A., & Muñoz, C. (2023). The effects of binge-watching and spacing on learning L2 multi-word units from captioned TV series. *The Language Learning Journal, 51*(4), 401–415. https://doi.org/10.1080/09571736.2023.2211614

Peters, E. (2019). The effect of imagery and on-screen text on foreign language vocabulary learning from audiovisual input. *TESOL Quarterly, 53*(4), 1008–1032. https://doi.org/10.1002/tesq.531

Peters, E., Heynen, E., & Puimège, E. (2016). Learning vocabulary through audiovisual input: The differential effect of L1 subtitles and captions. *System, 63*, 134–148. https://doi.org/10.1016/j.system.2016.10.002

Pu, P., Chang, D. Y.-S., & Wang, S. (2024). Incidental learning of collocations through different multimodal input: The role of learners' initial L2 proficiency. *System, 125*, 103416. https://doi.org/10.1016/j.system.2024.103416

Pu, P., Chang, D. Y.-S., & Wang, S. (2025). Incidental learning of multiword units from audiovisual input: The role of repetition and modality. *System, 131*, 103563. https://doi.org/10.1016/j.system.2025.103563

Puimège, E., & Peters, E. (2019). Learning formulaic sequences through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition, 42*(3), 525–549. https://doi.org/10.1017/S027226311900055X

Puimège, E., & Peters, E. (2020). Incidental learning of formulaic sequences from audiovisual input: Effects of learner- and item-related variables. *Studies in Second Language Acquisition, 42*(3), 525–549. https://doi.org/10.1017/S027226311900055X

Puimège, E., Montero Perez, M., & Peters, E. (2023). Promoting L2 acquisition of multiword units through textually enhanced audiovisual input: An eye-tracking study. *Second Language Research, 39*(2), 471–492. https://doi.org/10.1177/02676583211049741

Pujadas, G., & Muñoz, C. (2019). Extensive viewing of captioned and subtitled TV series: A study of L2 vocabulary learning by adolescents. *The Language Learning Journal, 47*(4), 479–496. https://doi.org/10.1080/09571736.2019.1616806

Radić-Bojanić, B. B. (2021). Audiovisual media and the acquisition of EFL vocabulary. *Nasleđe, 48*, 269–281. https://doi.org/10.46793/NasKg2148.269RB

Romero-Villamil, J. L., & Guzman-Martinez, C. P. (2020). Learning vocabulary through instructional subtitled videos. *Gist Education and Learning Research Journal, 21*, 7–25.

Sinyashina, E., & Balteiro, I. (2023). Incidental learning of word stress with captioned authentic videos: The effect of repetition. *Revista Electrónica de Lingüística Aplicada, 21*(1), 19–39. https://doi.org/10.58859/rael.v21i1.494

Soler Pardo, B. (2020). Subtitling and dubbing as teaching resources for learning English as a foreign language using Clipflair software. *Lenguaje y Textos, 51*, 41–56. https://doi.org/10.4995/lyt.2020.12690

Suárez, M. del M., & Gesa, F. (2019). Learning vocabulary with the support of sustained exposure to captioned video: Do proficiency and aptitude make a difference? *The Language Learning Journal, 47*(4), 497–517. https://doi.org/10.1080/09571736.2019.1617768

Suárez, M. del M., Gesa, F., & López, M. (2021). Vocabulary learning through sustained exposure to captioned video: Effects of proficiency and aptitude. *The Language Learning Journal, 49*(5), 623–640. https://doi.org/10.1080/09571736.2021.1876738

Teng, M. F. (2019a). Incidental vocabulary learning for primary school students: The effects of L2 caption type and word exposure frequency. *The Australian Educational Researcher, 46*, 113–136. https://doi.org/10.1007/s13384-018-0279-6

Teng, M. F. (2019b). The effects of video caption types and advance organizers on incidental L2 collocation learning. *Computers & Education, 142*, 103655. https://doi.org/10.1016/j.compedu.2019.103655

Teng, M. F. (2022a). Incidental L2 vocabulary learning from viewing captioned videos: Effects of learner-related factors. *System, 105*, 102736. https://doi.org/10.1016/j.system.2022.102736

Teng, M. F. (2022b). Vocabulary learning through videos: Captions, advance-organizer strategy, and their combination. *Computer Assisted Language Learning, 35*(3), 518–550. https://doi.org/10.1080/09588221.2020.1720253

Teng, M. F. (2023). The effectiveness of multimedia input on vocabulary learning and retention. *Innovation in Language Learning and Teaching, 17*(3), 738–754. https://doi.org/10.1080/17501229.2022.2131791

Teng, M. F. (2024). Incidental vocabulary learning from captioned video genres: Proficiency, working memory, and aptitude. *Computer Assisted Language Learning*, 1–43. https://doi.org/10.1080/09588221.2024.2421517

Teng, M. F. (2025a). Incidental vocabulary learning from captioned video genres: Vocabulary knowledge, comprehension, repetition, and working memory. *Computer Assisted Language Learning, 38*(5–6), 1301–1340. https://doi.org/10.1080/09588221.2023.2275158

Teng, M. F. (2025b). Modality of input and factors affecting incidental vocabulary learning: Reading, listening, and viewing with captions. *Applied Linguistics Review, 16*(4), 1607–1635. https://doi.org/10.1515/applirev-2024-0021

Teng, M. F., & Cui, Y. (2024). Comparing incidental learning of single words and collocations from different captioning conditions: The role of vocabulary knowledge and working memory. *Journal of Computer Assisted Learning, 40*(3), 973–989. https://doi.org/10.1111/jcal.12910

Teng, M. F., & Cui, Y. (2025a). Incidental vocabulary learning from captioned viewing: Relative contributions of word- and learner-related factors. *Language Teaching Research*. Advance online publication. https://doi.org/10.1177/13621688251372987

Teng, M. F., & Cui, Y. (2025b). Second language collocation learning through captioned videos: How do learners' vocabulary knowledge and working memory affect learning? *Computer Assisted Language Learning*. Advance online publication. https://doi.org/10.1080/09588221.2025.2497495

Teng, M. F., & Mizumoto, A. (2023). The role of spoken vocabulary knowledge in language minority students' incidental vocabulary learning from captioned television. *Australian Review of Applied Linguistics*, *46*(2), 113–136. https://doi.org/10.1075/aral.22033.ten

Teng, M. F., & Zhang, D. (2024). Vocabulary learning in a foreign language: Multimedia input, sentence-writing task, and their combination. *Applied Linguistics Review*, *15*(5), 2123–2148. https://doi.org/10.1515/applirev-2022-0160

van der Kolk, J., & Feijoo, S. (2024). Incidental vocabulary recognition effects of subtitled, captioned and reverse subtitled audiovisual input. *Revista de Lingüística y Lenguas Aplicadas, 19*, 218–230. https://doi.org/10.4995/rlyla.2024.18056

Vu, D. V., Noreillie, A.-S., & Peters, E. (2023). Incidental collocation learning from reading-while-listening and captioned TV viewing and predictors of learning gains. *Language Teaching Research*, 1–22. https://doi.org/10.1177/13621688221151048

Vulchanova, M., & Lervåg, I. K. (2021). Role of subtitles in L2 acquisition and comprehension: A pilot study of hearing-impaired students. *Languages*, *6*(1), Article 17. https://doi.org/10.3390/languages6010017

Vulchanova, M., Aurstad, L. M. G., Kvitnes, I. E. N., & Eshuis, H. (2015). As naturalistic as it gets: Subtitles in the English classroom in Norway. *Frontiers in Psychology*, *5*, Article 1510. https://doi.org/10.3389/fpsyg.2014.01510

Wang, A. (2025). The integration of auditory and textual input in vocabulary learning from subtitled viewing: An eye-tracking study. *Language Learning & Technology*, *29*(3), 70–91. https://doi.org/10.64152/10125/73648

Wang, A., & Pellicer-Sánchez, A. (2022). Incidental vocabulary learning from bilingual subtitled viewing: An eye-tracking study. *Language Learning*, *72*(3), 765–805. https://doi.org/10.1111/lang.12495

Wang, Y. (2019). Effects of L1/L2 captioned TV programs on students' vocabulary learning and comprehension. *CALICO Journal*, *36*(3), 204–224. https://doi.org/10.1558/cj.36268

Wi, I., & Boers, F. (2024). Sequential use of L1 and L2 captions: Exploring the benefits for vocabulary acquisition. *TESOL Quarterly*, *58*(1), 511–531. https://doi.org/10.1002/tesq.3243

Wu, H., & Yang, X. (2022). Effectiveness of textually enhanced captions on Chinese high-school EFL learners' incidental vocabulary learning. *Porta Linguarum*, *38*, 209–228. https://doi.org/10.30827/portalin.vi38.23511

Wu, H., Yu, P., Yang, S., & Chen, X. (2022). Video captioning effects on EFL listening comprehension and vocabulary learning. *International Journal of Computer-Assisted Language Learning and Teaching, 12*(2), 1–16. https://doi.org/10.4018/ijcallt.291534

Wu, W.-C. V., Lin, I.-T. D., Marek, M. W., & Ou Yang, F.-C. (2021). Analysis of English idiomatic learning behaviors of an audio-visual mobile application. *SAGE Open, 11*(2), 1–17. https://doi.org/10.1177/21582440211016899

Younas, M., & Dong, Y. (2024). The impact of using animated movies in learning English language vocabulary: An empirical study of Lahore, Pakistan. *SAGE Open, 14*(2), 1–12. https://doi.org/10.1177/21582440241258398

Yuan, X., & Tang, X. (2025). Effects of the sequential use of L1 and bilingual subtitles on incidental English vocabulary learning: A cognitive load perspective. *British Journal of Educational Psychology, 95*(2), 565–577. https://doi.org/10.1111/bjep.12740

Yüksel, D., & Tanrıverdi, B. (2009). Effects of watching captioned movie clips on vocabulary development of EFL learners. *The Turkish Online Journal of Educational Technology, 8*(2), Article 4.