









ARTICLE

Learner Perceptions of TikTok Video Features for Vocabulary Learning in ESL Contexts

Jane Xavierine ¹ , Norshima binti Zainal Shah ^{2*} , Mohd Hasrul bin Kamarulzaman ² , Joldasova Gozzal ³ ,
Jonathan Bryce ¹ , Saule Yeszhanova ⁴ , Kumutha Raman ¹ , Nur Izzati Khairuddin ¹ 

¹ Faculty of Education and Liberal Arts, INTI International University, Nilai 71800, Malaysia

² Language Centre, Universiti Pertahanan Nasional, Kuala Lumpur 57000, Malaysia

³ Department of Preschool and Primary Education, University of Innovation Technologies, Nukus, Karakalpakstan 230103, Uzbekistan

⁴ CARCEIT (Central Asian Research Centre for Educational Innovation and Transformation), Nazarbayev University Graduate School of Education, Astana 010000, Kazakhstan

ABSTRACT

Short-form video platforms such as TikTok have become central to how learners encounter English, yet their role in vocabulary development remains underexplored in TESOL research. Existing studies often describe TikTok as engaging but rarely identify which design features most effectively support vocabulary growth. This study addresses that gap by investigating ESL learners' perceptions of TikTok's multimodal features and by introducing the TikTok-Inspired Multimodal Vocabulary Framework (T-MVF). Ninety-two pre-intermediate learners at a Malaysian university watched curated TikTok videos that represented seven feature categories (subtitles, visuals, narrative, video format and duration, gestures, audio, and interactivity) before completing a 42-item bilingual questionnaire. Descriptive and correlational analyses revealed that subtitles, visuals, and narrative formed the strongest cluster, while gestures, audio, and interactivity were effective only when integrated with this core. Strong correlations confirmed that feature integration, rather than isolated use, underpins perceived vocabulary support. The study contributes empirically by identifying learner-valued feature clusters, theoretically by proposing the T-MVF to explain how design features

*CORRESPONDING AUTHOR:

Norshima binti Zainal Shah, Language Centre, Universiti Pertahanan Nasional, Kuala Lumpur 57000, Malaysia; Email: shima@upnm.edu.my

ARTICLE INFO

Received: 15 August 2025 | Revised: 22 September 2025 | Accepted: 25 September 2025 | Published Online: 9 December 2025

DOI: <https://doi.org/10.30564/fls.v7i12.11623>

CITATION

Xavierine, J., Zainal Shah, N.b., Kamarulzaman, M.H.b., 2025. Learner Perceptions of TikTok Video Features for Vocabulary Learning in ESL Contexts. *Forum for Linguistic Studies*. 7(12): 1826–1837. DOI: <https://doi.org/10.30564/fls.v7i12.11623>

COPYRIGHT

Copyright © 2025 by the author(s). Published by Bilingual Publishing Group. This is an open access article under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License (<https://creativecommons.org/licenses/by-nc/4.0/>).

moderate input and mediate outcomes, and pedagogically by offering evidence-based strategies for integrating short-form video into vocabulary instruction. The findings underscore the global potential of TikTok as a low-cost, widely accessible, and engaging tool for vocabulary development. By clarifying which features matter most, the study advances multimodality research, informs ESL and EFL classroom practice, and aligns with Sustainable Development Goal 4 on Quality Education.

Keywords: TikTok; Vocabulary Learning; English as a Second Language; Multimodality; Captions; Short Form Video; Instructional Design; Quality Education

1. Introduction

1.1. Background and Problem

Short form video platforms have changed how young people access and process information. TikTok in particular is widely used not only for entertainment but also for informal learning. TikTok, YouTube Shorts, and Instagram Reels provide similar short clips, but TikTok stands out for its algorithm and ease of content creation. These characteristics make it especially powerful as an informal learning environment. The platform's reach is also unprecedented, with numerous clips shared daily across diverse cultures, giving learners frequent, authentic exposure to new words and expressions. Unlike traditional classroom materials, TikTok provides immediacy, variety, and opportunities for incidental encounters with language in everyday contexts. Many students report that they first encounter English expressions on TikTok even when they are not actively studying. This spontaneous exposure highlights the platform's potential for vocabulary development.

TikTok integrates speech, text overlays, images, gestures, music, and visual effects in short clips, and such platforms are described as part of the broader digital transformation reshaping education, work, and communication during the COVID-19 era ^[1]. This transformation is also reflected in shifting learning preferences within higher education, where students increasingly rely on digital tools for both formal and informal study ^[2]. These environments are described as multimodal, with meaning produced through the interaction of linguistic, visual, gestural, spatial, and audio modes ^[3].

Vocabulary is central to second language learning. Without it, learners cannot read fluently, express ideas clearly, or understand spoken input. While vocabulary learning requires both breadth and depth ^[4], repeated encounters across contexts are proven to support lasting

retention ^[5,6]. Lists often fade after a test, but words encountered through multiple channels and contexts are remembered longer.

Outside the classroom, vocabulary growth is increasingly supported by digital environments. Research on extramural English shows that informal contact with media contributes significantly to language development ^[7]. TikTok may serve as such an environment, offering authentic, engaging, and repeated encounters with vocabulary.

1.2. Theoretical Foundations

TikTok's design aligns with several influential theories of learning. Multimodality theory argues that meaning emerges from the interplay of multiple modes, such as linguistic, visual, gestural, and audio ^[3]. Dual coding theory explains that retention is enhanced when learners process information through both verbal and visual systems ^[8]. The multimedia principles highlight the benefits of synchronising captions and images (temporal contiguity), reducing irrelevant distractions (coherence), and reinforcing input through repetition (redundancy) ^[9]. Cognitive load theory further emphasises that when multiple modes are combined, design must remain simple, clear, and well-paced so that learners' attention is directed to essential content rather than split across excessive or competing stimuli ^[10]. Together, these theories suggest that learners may find captions, visuals, and short narratives on TikTok particularly valuable for vocabulary learning.

Recent studies provide further support, confirming that captions and subtitles assist learners in noticing and recalling vocabulary ^[11,12]. Not only are visuals proven to promote memory through mental imagery, but research also confirms that presenting vocabulary through combined verbal and visual cues strengthens retention. Recent work shows that multimedia annotations significantly support learners in constructing mental images and processing new vocabulary

more deeply, especially when visual cues are synchronised with textual explanations^[12], but it is also argued that narrative makes words more memorable by embedding them in meaningful contexts^[13]. These findings align closely with the affordances of TikTok, where captions, visuals, and narratives are often combined in short clips.

1.3. Research Gaps

Despite these promising links, empirical research on TikTok in ESL remains limited. Pas studies found that learners perceived TikTok as useful for vocabulary learning^[14], while TikTok microlearning environments are reported to have generated high engagement compared to more traditional tools^[15]. Students have noted that they view TikTok as a valuable resource for both English vocabulary and pronunciation practice^[16,17]. Other study by Abdelrahman and Salama^[18] also examined TikTok in relation to motivation and found that TikTok's mobile-assisted environment encouraged positive engagement and offered practical opportunities for vocabulary and pronunciation practice. Their findings highlight the platform's potential to support learning, although they did not identify which specific design features contribute most to these outcomes.

This absence of feature-level analysis leaves teachers without clear guidance on how to use TikTok in pedagogical contexts. Moreover, existing research often stops at measuring engagement or motivation rather than learning outcomes. Teachers therefore face the challenge of making context-sensitive decisions without evidence about which specific TikTok features best support vocabulary learning. This problem is not only relevant in Malaysia but also in other regions where learners are heavy users of TikTok,

including East Asia, the Middle East, and Europe. Understanding the impact of TikTok's design features has global significance, since short-form video is accessible, low-cost, and widely adopted.

1.4. The TikTok-Inspired Multimodal Vocabulary Framework (T-MVF)

To address this gap, this study introduces the TikTok-Inspired Multimodal Vocabulary Framework (T-MVF) (Figure 1). The framework integrates multimodality, dual coding, and multimedia learning principles to conceptualise how TikTok features support vocabulary development. It consists of four interconnected components.

Multimodal Input serves as the independent variable and includes linguistic, visual, gestural, audio, spatial, and interactive resources. TikTok Design Features act as a moderator, shaping how learners process input through subtitles, video duration, visuals, narrative, gestures, and interactivity. Enhancement of Vocabulary functions as the mediator, capturing the cognitive and affective processes of noticing, contextualising, and rehearsing new words. Finally, Vocabulary Learning Outcomes serve as the dependent variable, representing measurable gains in comprehension, recall, and productive use of vocabulary.

By positioning enhancement as the mediator and outcomes as the dependent variable, the T-MVF clarifies that vocabulary learning depends not only on exposure but also on how design features moderate input and how enhancement processes mediate this relationship. This structure provides a logical pathway that explains not only whether TikTok supports vocabulary learning, but also how and why it does so.

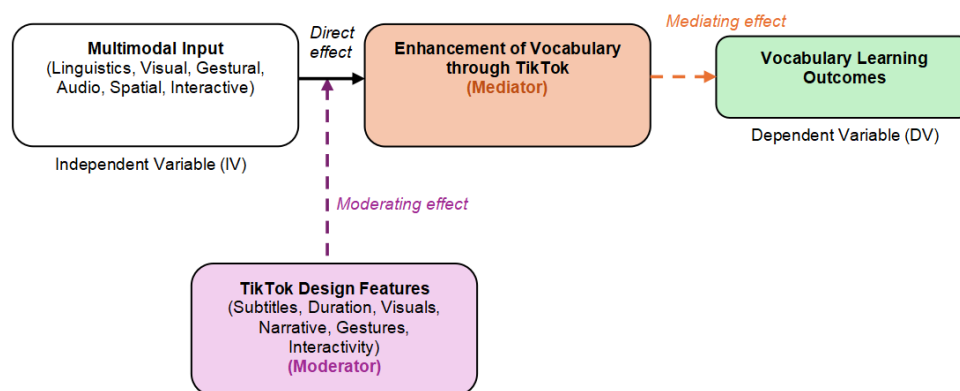


Figure 1. TikTok-Inspired Multimodal Vocabulary Framework (T-MVF).

1.5. Research Aim and Contribution

This study investigates learner perceptions of seven TikTok feature categories: subtitles and text, visuals, narrative, video format and duration, gestures, audio, and interactivity. Ninety-two pre-intermediate learners in Malaysia completed a bilingual questionnaire after viewing curated TikTok videos. The study aims to identify which features learners value most and how these features interact to support vocabulary development.

The contribution of this study is threefold. Empirically, it identifies the integrated cluster of captions, visuals, and narrative as the most effective for vocabulary learning. Theoretically, it proposes the T-MVF as a new framework for explaining how multimodal input, design features, and mediating processes lead to vocabulary outcomes. Pedagogically, it offers evidence-based strategies for teachers to select or design TikTok content that supports vocabulary learning. This contribution aligns with Sustainable Development Goal 4 (Quality Education) by promoting innovative, inclusive, and accessible approaches to language education across diverse contexts.

2. Materials and Methods

2.1. Research Design

This study employed a quantitative survey design to examine how ESL learners perceive different TikTok features in relation to vocabulary learning. A survey method was chosen because it enabled the efficient collection of standardised responses across a large group, making it possible to conduct both descriptive analysis of feature preferences and correlational analysis of inter-feature relationships^[19].

2.2. Participants

Ninety-two students enrolled in the Intensive English Programme (IEP) at a private Malaysian university participated. Placement test results classified all students at pre-intermediate level (CEFR A2). Most participants were between 18 and 20 years old, with some slightly older. The group was culturally diverse, with students from across Asia and the Middle East, creating a representative context

for exploring how learners with varied educational backgrounds respond to TikTok's multimodal features.

Within this cohort, most were recent secondary school graduates, while a smaller number were older students returning to strengthen their English for academic or professional reasons. About half reported using TikTok daily, while others engaged with the platform only occasionally or preferred alternatives such as Instagram or YouTube. This variation allowed comparisons across different levels of familiarity with short-form video. Gender distribution was balanced, and students represented regions including Malaysia, China, Saudi Arabia, and Central Asia. Such diversity provided a strong foundation for examining shared as well as distinctive perceptions of TikTok's multimodal features.

Participation was voluntary and unrelated to course assessment. Informed consent was obtained, and ethical approval was granted by the Universiti Pertahanan Nasional Malaysia Research Ethics Committee (Protocol code: UPNM (PBH) 16.02/05 (4)).

2.3. Instrument Development and Validation

The instrument for this study was a 42-item bilingual questionnaire, available in English and Mandarin to ensure accessibility for pre-intermediate learners, the majority of whom were Chinese. The instrument was designed to measure perceptions of seven categories of TikTok features: (1) Subtitles and Text, (2) Visual Elements, (3) Narrative and Story, (4) Video Format and Duration, (5) Gestural and Spatial Elements, (6) Audio Elements, and (7) Interactive Elements. Each category contained six items, rated on a five-point Likert scale ranging from strongly disagree (1) to strongly agree (5).

The development process followed three stages to ensure rigor and construct validity. First, an initial item pool was generated through an extensive review of literature on multimodal learning, vocabulary acquisition, and short-form video design^[9,11,12,20]. Recent reviews of mobile-assisted vocabulary learning highlight that digital tools increasingly integrate multimodal input and visual-textual combinations, reinforcing the need for instruments that reflect these contemporary learning environments^[20]. Second, the draft questionnaire underwent expert validation using the Delphi method. Five specialists in TESOL, edu-

cational technology, multimodality, and multimedia learning reviewed the items in two iterative rounds. Their feedback confirmed content validity and improved accessibility for learners. For example, the item “Subtitles assist me in constructing connections between form and meaning” was revised to the simpler “Captions help me understand the meaning of new words.” This ensured that the language of the instrument was appropriate for A2 learners.

Third, a pilot study was conducted with a separate group of 30 pre-intermediate students from the same programme. Pilot data were analysed for internal consistency. Results showed a Cronbach’s alpha of 0.94 for the overall scale, which exceeds the conventional 0.80 threshold in social science research. Reliability was also high across all categories, with alpha values ranging from 0.86 to 0.91. These values indicate strong internal consistency and support the use of the instrument for the main study.

The final instrument therefore combined theoretical grounding, expert validation, and empirical reliability testing. It provided a tool that was both statistically sound and accessible to learners, ensuring that responses captured genuine perceptions of TikTok’s multimodal features without being hindered by linguistic complexity.

2.4. Procedure

The study was conducted in a classroom setting to ensure equal exposure. Seven TikTok videos were selected to represent the seven feature categories. Each video lasted under 60 seconds, contained standard English pronunciation, and focused on one dominant feature. Videos with distracting edits, slang, or culturally inappropriate material were excluded to maintain focus and consistency.

Each video was shown twice. The first viewing allowed learners to grasp the overall message, while the second encouraged closer attention to specific features. This approach aligns with the redundancy principle^[9], which highlights the value of repeated multimodal input. The curated set included, for example, a clip with prominent captions, another embedding vocabulary in a short story, and one that used exaggerated gestures to illustrate meaning.

After the viewing session, learners completed the questionnaire digitally on their own devices. Online administration mirrored the way learners normally engage

with TikTok and therefore enhanced ecological validity. On average, completion time was 20 minutes. Learners commented that the bilingual format improved clarity and confidence in their responses.

2.5. Data Analysis

Data was analysed using IBM SPSS Statistics version 27. Descriptive statistics were computed to identify mean scores and standard deviations for each feature category, providing an overview of learners’ perceptions. Pearson product–moment correlation coefficients were then calculated to examine relationships among categories. Statistical significance was set at $p < 0.01$ to reduce the risk of Type I error.

This combination of descriptive and correlational analysis offered both breadth and depth. Descriptive results revealed general patterns of learner preference, while correlations highlighted how features interacted. For instance, captions might be valued on their own but could also prove more effective when paired with visuals. Although correlations are useful for identifying associations, they cannot determine predictive strength or causal direction. Future research should therefore apply structural equation modelling (SEM) to test the TikTok-Inspired Multimodal Vocabulary Framework (T-MVF) empirically, modelling both direct and indirect effects among the independent, moderator, mediator, and dependent variables.

2.6. Ethical Considerations

The study was designed with strict attention to ethical standards. Participation was voluntary, and learners were assured that their responses would remain anonymous and confidential. They were also informed of their right to withdraw at any time without penalty. In addition, informed consent was obtained prior to data collection.

The Research Ethics Committee of Universiti Pertahanan Nasional Malaysia approved the study (protocol code: UPNM (PBH) 16.02/05 (4)). The use of TikTok videos complied with the platform’s terms of service, and all selected videos were screened to ensure they were culturally appropriate and suitable for pre-intermediate learners.

3. Results

3.1. Descriptive Analysis

Learners responded positively to all seven categories of TikTok features, with every category scoring above the high agreement threshold of 3.4. Subtitles and Text received the highest mean score ($M = 3.92$, $SD = 0.86$),

followed closely by Visual Elements ($M = 3.90$, $SD = 0.84$) and Narrative and Story ($M = 3.84$, $SD = 0.91$). The remaining categories, Video Format and Duration ($M = 3.78$, $SD = 0.79$), Gestural and Spatial Elements ($M = 3.70$, $SD = 0.88$), Audio Elements ($M = 3.70$, $SD = 0.89$), and Interactive Elements ($M = 3.67$, $SD = 0.90$) which also received favourable ratings, though at slightly lower levels (Table 1).

Table 1. Descriptive statistics for learner perceptions of TikTok features for vocabulary learning.

Feature Category	Mean	SD
Subtitles and Text	3.92	0.86
Visual Elements	3.90	0.84
Narrative and Story	3.84	0.91
Video Format and Duration	3.78	0.79
Gestural and Spatial Elements	3.70	0.88
Audio Elements	3.70	0.89
Interactive Elements	3.67	0.90

The preference for subtitles confirmed previous research on captioning^[10,11,21]. Subtitles allow learners to read and hear words simultaneously, strengthening the link between form and meaning. This behaviour aligns with Schmidt's noticing hypothesis^[22], which emphasises that learners must consciously attend to linguistic features for acquisition to occur. Many participants explained in informal comments that captions gave them confidence in recognising words they had only partially understood through listening.

Visual clarity in digital learning environments has been emphasised in recent studies, which show that well-designed visual supports and multimedia cues contribute to improved vocabulary retention^[12,23]. Meta-analytic evidence indicates that technology-assisted tools are most effective when visual input is clear, well-paced, and aligned with learners' processing needs^[23]. Learners in this study frequently commented that they could "picture" a word when recalling it later, suggesting that visual association was a strong support mechanism.

Narrative and Story emerged as another highly valued feature. Words embedded in short stories were easier to recall and more meaningful^[13]. Learners responded positively to videos that presented vocabulary in relatable contexts, such as short skits or daily-life situations. Several participants noted that narrative made the vocabulary feel

"real" rather than abstract. The role of narrative parallels findings from music-based learning, where repeated exposure to words in songs supports incidental vocabulary growth^[24]. In TikTok, micro-narratives within short clips function in a similar way, sustaining engagement while introducing new words.

Secondary features also contributed, although at slightly lower levels. Gestures, for example, reinforced meaning when paired with intonation. Learners noticed when a teacher or presenter exaggerated a movement to emphasise the meaning of a word such as "tiny" or "huge." This echoes research showing that gestures are not mere add-ons but integral to thought and communication^[25]. While gestures ranked below captions and visuals, their supportive role was clear, particularly when synchronised with speech^[26,27].

Audio was rated at a similar level to gestures. Students valued clear pronunciation, intonation, and rhythm, especially when paired with subtitles. However, some participants commented that music or background noise occasionally distracted from the spoken word. This suggests that audio is effective when controlled carefully but may hinder learning if overused.

Interactive features, such as polls or on-screen prompts, received the lowest score, though still above neutral. Learners often felt that such features distracted from

processing the vocabulary. Many preferred interactivity after viewing rather than during the clip, as pop-ups or rapid questions could be overwhelming. This preference may reflect findings from recent reviews showing that mobile-assisted vocabulary learning environments are most effective when multimodal input is paced clearly and without unnecessary cognitive burden^[20].

Taken together, these findings suggest that subtitles, visuals, and narrative form a “core cluster” of features

most directly linked to vocabulary support, while gestures, audio, and interactivity play supporting roles when designed appropriately.

3.2. Correlation Analysis

Pearson correlation coefficients were calculated to explore relationships among the seven feature categories (Table 2).

Table 2. Pearson correlations among TikTok feature categories.

Feature Pair	r	p-Value	Interpretation
Narrative and Story with Visual Elements	0.868	<0.01	Very strong positive
Subtitles and Text with Format and Duration	0.853	<0.01	Very strong positive
Visual Elements with Subtitles and Text	0.841	<0.01	Very strong positive
Interactive Elements with Format and Duration	0.805	<0.01	Strong positive
Interactive Elements with Subtitles and Text	0.804	<0.01	Strong positive
Audio Elements with Gestural and Spatial Elements	0.758	<0.01	Strong positive

The strongest correlation was between Narrative and Visual Elements ($r = 0.868$). Learners perceived stories and images as inseparable, suggesting that narrative contexts became more memorable when reinforced visually. This supports earlier findings that narrative enhances retention through mental imagery^[13]. Learners in this study often described remembering a word when they could recall the scene or action in which it was used, underscoring the synergy of narrative and visual input. Broader evidence from technology-assisted vocabulary learning research further suggests that digital tools that combine narrative, visual clarity, and multimodal cues tend to produce stronger vocabulary outcomes^[23].

A similarly strong correlation was observed between Subtitles and Video Format and Duration ($r = 0.853$). Learners valued captions most when the clip was long enough for comfortable processing. Videos that were too fast reduced the usefulness of subtitles, while well-paced clips allowed time to read, listen, and connect meaning. This finding echoes earlier research demonstrating that caption effectiveness depends on processing time^[11,21].

The correlation between Visual Elements and Subtitles ($r = 0.841$) further highlighted the synergy between text and imagery. Learners recalled words more successfully when captions were paired with clear visuals, reinforcing dual coding theory^[8]. Recent evidence from multime-

dia annotation studies confirms that integrated visual–text combinations facilitate vocabulary noticing and retention by providing mental imagery support^[12].

Interactive Elements correlated strongly with both Subtitles ($r = 0.804$) and Format and Duration ($r = 0.805$). This indicates that interactivity was most effective when tied to well-paced, captioned videos. Learners were more receptive to answering a prompt or engaging with vocabulary when they had enough time to process the content. Poorly timed interactivity, however, could create cognitive overload. This finding aligns with cognitive load theory^[9] and with recent studies on mobile-assisted vocabulary learning^[20].

Finally, Audio Elements and Gestural Elements correlated strongly ($r = 0.758$). Learners recognised how sound and movement reinforced each other. For example, exaggerated intonation paired with a gesture, such as miming “gigantic,” strengthened understanding. This resonates with research showing that gesture–speech synchrony enhances recall^[26,27].

3.3. Summary of Results

Overall, the results show a consistent pattern. Learners placed the highest value on subtitles, visuals, and narrative, and these features also demonstrated the strongest

interrelationships. The descriptive data confirmed their importance, while the correlation analysis showed that these features function best when integrated. Secondary features such as gestures, audio, and interactivity were also valued, but their contribution was contingent on alignment with the core cluster.

The results therefore suggest that effective vocabulary learning through TikTok depends less on isolated features and more on the integration of multimodal resources. When subtitles, visuals, and narratives work together, learners are better able to notice, understand, and retain new vocabulary. Supporting features can enhance this process, but only when they align with duration, clarity, and pacing.

4. Discussion

4.1. Empirical Contribution

This study provides empirical evidence on how ESL learners perceive different TikTok features in relation to vocabulary learning. The descriptive findings showed that subtitles, visuals, and narrative were rated most positively. The correlation analysis confirmed that these features were also most effective in combination rather than isolation. Together, they form a “core cluster” that learners consistently rely on when processing new vocabulary. Subtitles supported noticing and form–meaning mapping, visuals reinforced memory through mental imagery, and narratives provided meaningful contexts that made vocabulary easier to recall.

These findings align with and extend prior research on multimedia learning. Previous work on captioning established that subtitles enhance listening comprehension and incidental vocabulary growth^[10,11,21]. The present study adds that learners themselves report captions as more useful when integrated with clear visuals and narratives. Similarly, research on multimedia-enhanced vocabulary learning has shown that visuals paired with text promote stronger retention by enabling learners to generate mental images of new words^[12,23]. Our results support this claim but also reveal that visuals are most effective when combined with narrative contexts. This provides new evidence that integration across modes amplifies learning outcomes.

When compared with research on long-form media

such as television and YouTube, the present findings highlight the unique role of brevity in short-form video. Learners valued subtitles and visuals most when clips were short enough to process comfortably. Longer or faster-paced videos risked overwhelming their attention, reducing the usefulness of captions. This confirms that pacing and duration are not peripheral concerns but essential design features that influence vocabulary uptake. Teachers who curate or produce TikTok content should therefore prioritise brevity and clarity alongside multimodal integration.

Secondary features such as gestures, audio, and interactivity also contributed but in conditional ways. Gestures were effective when paired with intonation, echoing earlier findings that gesture–speech synchrony enhances recall^[26,27]. Audio was valued when clear and free from distracting background noise, while interactivity and duration worked best when carefully timed. These results suggest that supporting features play a scaffolding role rather than a central one. This challenges assumptions that adding multiple modes automatically improves learning. Instead, effectiveness depends on careful alignment with the core cluster of captions, visuals, and narrative.

4.2. Theoretical Contribution

The findings also refine and extend theoretical perspectives. The TikTok-Inspired Multimodal Vocabulary Framework (T-MVF) positions multimodal input as the independent variable, TikTok design features as the moderator, enhancement of vocabulary as the mediator, and vocabulary learning outcomes as the dependent variable. By showing that learners perceive clusters of features as more effective than isolated ones, the results validate the framework’s emphasis on integration.

The strong correlations among subtitles, visuals, and narrative extend Paivio’s dual coding theory^[8] by demonstrating that verbal and non-verbal input is most powerful when aligned in time and context. They also reinforce Mayer’s multimedia principles^[9], particularly temporal contiguity (captions and visuals appearing together), coherence (removal of irrelevant elements), and redundancy (repetition across modes). Learners’ consistent preference for short, well-paced videos further demonstrates that design features moderate whether multimodal input facilitates or hinders processing.

From a multimodality perspective ^[3,28], the study confirms that meaning emerges not from single modes but from the orchestration of several semiotic resources. TikTok provides a unique case because it compresses speech, text, visuals, and gesture into clips lasting less than a minute. Learners in this study recognised that meaning was shaped by the integration of these resources, not by their individual presence. In this way, the T-MVF contributes a new model for understanding how short-form video mediates vocabulary learning through multimodal integration.

At the same time, the study highlights the need for more advanced modelling. Correlations can reveal associations but cannot establish predictive pathways. Future research should apply structural equation modelling (SEM) to test the T-MVF empirically. Such analysis could measure the direct and indirect effects among multimodal input, design features, vocabulary enhancement, and learning outcomes. Confirming these pathways would strengthen the framework's theoretical contribution and position it as a robust model for future studies on digital multimodality and vocabulary learning.

4.3. Pedagogical Contribution

The study also carries significant pedagogical implications. Since subtitles, visuals, and narrative emerged as the most valued features, teachers should prioritise clips that combine these elements. For example, a video introducing the word “fragile” might show the word in captions, present an image of a breaking glass, and embed it in a short story. This integration maximises noticing, reinforces mental imagery, and situates the word in meaningful use.

Secondary features should be treated as scaffolds. Gestures and intonation can enhance meaning when aligned carefully with speech, while interactive prompts can encourage learners to practise using new vocabulary. However, poorly designed features may distract rather than support. For example, quizzes embedded in fast-paced clips risk overloading learners. Teachers therefore need to evaluate TikTok content critically, focusing on design quality rather than assuming that more multimodality automatically leads to better outcomes.

In practice, TikTok can be integrated into classroom instruction through short cycles of input, noticing, and

practice. Learners might first watch a curated clip, then complete tasks such as identifying target words in captions or matching visuals with meanings. This should be followed by productive activities such as retelling the story, using the words in new sentences, or creating original short clips. Learner-generated TikTok content not only reinforces vocabulary but also promotes creativity, digital literacy, and confidence in using English.

Beyond the classroom, these implications extend to global contexts. TikTok is widely used across Asia, Europe, and the Middle East, making it a scalable and accessible resource. Policymakers and curriculum designers can consider short-form video as a low-cost strategy for supporting vocabulary learning in diverse educational systems. The findings also align with Sustainable Development Goal 4 on Quality Education by promoting inclusive and innovative approaches that can reach learners across different socioeconomic backgrounds

4.4. Theoretical Significance and Limitations

While the study provides valuable insights, several limitations must be acknowledged. The reliance on descriptive and correlational data restricts claims about causality. Learners' perceptions, though useful, may not directly reflect measurable vocabulary gains. Self-report data also risk social desirability bias, as learners may respond positively to appear engaged.

Future research should combine perception surveys with objective measures such as vocabulary pre- and post-tests, eye-tracking studies, or recall tasks to capture real-time processing. Longitudinal designs would reveal whether vocabulary learned through TikTok is retained over time. Applying SEM would allow testing of the T-MVF as an integrated model, validating its pathways across different contexts. Expanding to advanced learners or younger school-aged learners would also strengthen the generalisability of the framework.

Another avenue for future research involves examining how artificial intelligence and personalised algorithms shape exposure. TikTok's recommendation system may reinforce repeated encounters with certain words, potentially supporting retention, but it may also limit learners to narrow linguistic inputs. Exploring this dynamic would extend understanding of how digital ecosystems influence

language acquisition.

5. Conclusions

This study examined how ESL learners perceive TikTok's multimodal features for vocabulary learning. The findings revealed that subtitles, visuals, and narrative form the strongest cluster of features, while gestures, audio, and interactivity play a supportive role when integrated effectively. These results highlight that vocabulary learning is shaped not by isolated features but by the interaction of multimodal input, design features, and mediating processes.

The study contributes empirically by identifying feature clusters that learners value most, theoretically by refining the TikTok-Inspired Multimodal Vocabulary Framework (T-MVF), and pedagogically by offering practical strategies for teachers to integrate short-form video into instruction. By positioning Enhancement of Vocabulary as the mediator and Vocabulary Learning Outcomes as the dependent variable, the T-MVF clarifies how exposure to multimodal input is transformed into measurable gains.

These contributions extend beyond the Malaysian context. TikTok and similar short-form video platforms are widely accessible worldwide, making the framework relevant to diverse ESL and EFL settings. Teachers in different regions can draw on these findings to make informed decisions about which features to prioritise in their classrooms. Policymakers and curriculum designers can also use the results to recognise short-form video as a cost-effective and engaging tool aligned with Sustainable Development Goal 4 on Quality Education. Importantly, the findings highlight how inclusive, accessible, and low-cost technologies can expand equitable access to quality language learning, particularly for learners with limited resources and in line with important global SDG 4 targets.

The study has limitations, particularly the reliance on descriptive and correlational analyses. While these revealed clear associations, they do not confirm predictive strength or causal direction. Future research should therefore employ structural equation modelling (SEM) to validate the T-MVF as an integrated model and to test its pathways statistically. Longitudinal studies that track vocabulary retention and cross-contextual research in differ-

ent countries and proficiency levels would further strengthen its applicability.

In conclusion, this study demonstrates that TikTok can be more than a source of entertainment. When captions, visuals, and narratives are integrated thoughtfully, and when supportive features such as gestures and interactivity are aligned, short-form video becomes a powerful tool for vocabulary development. The T-MVF provides a framework for understanding how this process unfolds and offers a guide for teachers and researchers seeking to harness popular media for meaningful language learning. Looking ahead, future work can explore how emerging technologies, such as AI-generated captions, adaptive content, or personalised recommendation systems, might further enhance the effectiveness of short-form video for vocabulary acquisition. By bridging theory, empirical evidence, and classroom practice, the study contributes to the growing recognition that digital platforms can transform informal exposure into sustained language learning outcomes.

Author Contributions

Conceptualization, J.X.; methodology, J.X., N.b.Z.S. and N.I.K.; validation, M.H.b.K., K.R. and J.B.; formal analysis, J.X.; investigation, J.X. and J.G.; resources, J.B. and K.R.; data curation, J.X. and N.b.Z.S.; software, N.I.K.; writing—original draft preparation, J.X.; writing—review and editing, N.b.Z.S., M.H.b.K., J.G., J.B., S.Y., K.R. and N.I.K.; visualization, J.X. and N.b.Z.S.; supervision and project administration, N.b.Z.S. and M.H.b.K. All authors have read and agreed to the published version of the manuscript.

Funding

This work received no external funding.

Institutional Review Board Statement

The study was conducted in accordance with the Declaration of Helsinki and approved by the Research Ethics Committee of Universiti Pertahanan Nasional Malaysia

(protocol code UPNM (PBH) 16.02/05 (4), approved on 12 June 2025).

Informed Consent Statement

Informed consent was obtained from all subjects involved in the study.

Data Availability Statement

The data supporting the findings of this study are not publicly available due to ethical restrictions and the need to protect participant confidentiality. Data may be made available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest.

References

- [1] Dwivedi, Y.K., Hughes, L., Coombs, C., et al., 2020. Impact of COVID-19 pandemic on information management research and practice: Transforming education, work and life. *International Journal of Information Management*. 55, 102211. DOI: <https://doi.org/10.1016/j.ijinfomgt.2020.102211>
- [2] Sagadavan, R., John, S., 2019. Learning preferences transformation in tertiary education. *International Journal of Recent Technology and Engineering (IJRTE)*. 8(2S), 215–220.
- [3] New London Group, 1996. A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review*. 66(1), 60–92.
- [4] Nation, I.S.P., 2013. *Learning vocabulary in another language*, 2nd ed. Cambridge University Press: Cambridge, UK.
- [5] Webb, S., 2008. Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition*. 30(1), 79–95.
- [6] Webb, S., Nation, I.S.P., 2017. *How vocabulary is learned*. Oxford University Press: Oxford, UK.
- [7] Sundqvist, P., Sylén, L.K., 2016. *Extramural English in teaching and learning: From theory and research to practice*. Palgrave Macmillan: London, UK.
- [8] Paivio, A., 1986. *Mental representations: A dual coding approach*. Oxford University Press: Oxford, UK.
- [9] Johnson, C.I., Mayer, R.E., 2009. A testing effect with multimedia learning. *Journal of Educational Psychology*. 101(3), 621–629.
- [10] Sweller, J., 1998. Cognitive load during problem solving: Effects on learning. *Cognitive Science*. 12(2), 257–285.
- [11] Montero Pérez, M., Peters, E., Clarebout, G., et al., 2014. Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology*. 18(1), 118–141.
- [12] Zou, D., Teng, M.F., 2023. Effects of tasks and multimedia annotations on vocabulary learning. *System*. 115, 103050.
- [13] Boers, F., Lindstromberg, S., 2012. Experimental and intervention studies on formulaic sequences in a second language. *Annual Review of Applied Linguistics*. 32, 83–110.
- [14] Alharthy, F., 2025. Exploring the impact of TikTok on second-language vocabulary acquisition: Benefits, challenges, and learner perceptions. *Journal of Humanities and Social Sciences Studies*. 7(3), 22–31.
- [15] Conde-Caballero, D., Murillo-Pérez, M.D., Díaz-Méndez, C., 2023. Microlearning through TikTok in higher education: An evaluation of uses and potentials. *Education and Information Technologies*. 28(11), 11977–11999.
- [16] Fitria, T.N., 2023. Value engagement of TikTok: A review of TikTok as learning media for language learners in pronunciation skill. *EBONY: Journal of English Language Teaching, Linguistics, and Literature*. 3(2), 91–108.
- [17] Mei, B., Aziz, A., 2022. Students' perception on using TikTok application as an English learning tool. *International Journal of Academic Research in Progressive Education and Development*. 11(4), 202–213.
- [18] Zaghar, F., 2022. Higher Education in Times of Pandemic: An Exploration of Teachers and EFL Learners' Perceptions of the Shift to Online Instruction Option. *Arab World English Journal (AWEJ)*. 8, 205–213.
- [19] Creswell, J.W., Creswell, J.D., 2017. *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications: Thousand Oaks, CA, USA.
- [20] Lin, Z., Lin, Y., 2025. A review of research on mobile-assisted vocabulary learning. *International Journal of Emerging Technologies in Learning*. 20(3), 1–15.
- [21] Wei, R., Teng, F., Zhou, S., 2022. On-screen texts in audiovisual input for L2 vocabulary learning: A review. *Frontiers in Psychology*. 13, 904523.
- [22] Schmidt, R., 1990. The role of consciousness in sec-

- ond language learning. *Applied Linguistics*. 11(2), 129–158.
- [23] Hao, T., Wang, Z., Ardasheva, Y., 2021. Technology-assisted vocabulary learning for EFL learners: A meta-analysis. *Journal of Research on Educational Effectiveness*. 14(3), 645–667.
- [24] Peters, E., Webb, S., 2018. Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*. 40(3), 551–577.
- [25] McNeill, D., 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago Press: Chicago, IL, USA.
- [26] Nirme, J., Gulz, A., Haake, M., et al., 2024. Early or synchronized gestures facilitate speech recall—a study based on motion capture data. *Frontiers in Psychology*. 15, 1345906.
- [27] Macedonia, M., 2025. Your body as a tool to learn second language vocabulary. *Behavioral Sciences*. 15(8), 997.
- [28] Kress, G., 2010. *Multimodality: A social semiotic approach to contemporary communication*. Routledge: London, UK.