JOURNAL OF COMMUNICATION, LANGUAGE AND CULTURE

Multimodal Orchestration in TikTok Vocabulary Videos: A Qualitative Analysis of Vocabulary Meaning Making

Jane Xavierine[1*], Norshima binti Zainal Shah[2], Mohd Hasrul bin Kamarulzaman[2], Alice Shanthi[3], Jonathan Bryce[1]

[1]Faculty of Education and Liberal Arts, INTI International University, Nilai, Malaysia
[2]Language Centre, Universiti Pertahanan Nasional Malaysia, Kuala Lumpur, Malaysia
[3]Akademi Pengajian Bahasa, Universiti Teknologi MARA (UiTM) Cawangan Negeri Sembilan, Seremban, Malaysia

*Corresponding author: jane.xavier@newinti.edu.my; ORCID iD: 0000-0003-1296-9516

ABSTRACT

The use of TikTok for short-form vocabulary instruction has become increasingly prevalent; however, empirical work examining the instructional design of these videos, particularly how multimodal elements are combined to construct vocabulary meaning, remains limited. This study uses qualitative multimodal discourse analysis to investigate how vocabulary meaning is constructed across ten English vocabulary videos (five institutional and five creator-driven). A theoretically informed coding frame was developed to analyse five semiotic modes (linguistic, visual, auditory, embodied, and spatial) through timestamped co-occurrence analysis in ATLAS.ti. The findings suggest that the meaning of vocabulary in these videos is represented through the coordinated orchestration of multiple modes rather than through isolated modal inputs. Patterns of co-occurrence point to recurring configurations consistent with signalling, redundancy, and attentional guidance. Temporal density, defined as the concentration of multimodal cues within brief instructional intervals, emerged as a salient design feature. Creator-driven videos exhibited higher multimodal density and performance convergence than institutional videos. This study contributes to multimodality research by foregrounding temporal orchestration as an important analytic dimension in short-form instructional videos.

Keywords: multimodal discourse analysis, TikTok, vocabulary instruction, short-form video, temporal density

Received: 13 January 2026, **Accepted:** 2 April 2026, **Published:** 30 July 2026

doi: https://doi.org/10.33093/jclc.2026.6.2.17

Published by MMU Press. URL: https://journals.mmupress.com/jclc

1.0 Introduction

TikTok has emerged as a prominent platform through which learners encounter short-form vocabulary content (Waroh et al., 2025; Zeng et al., 2021). Originally intended for short-form entertaining media, the platform has recently emerged as a place to find educational content about language, specifically vocabulary explanations meant to be learned quickly and applied immediately. TikTok vocabulary videos combine an array of multimodal elements – spoken explanation, on-screen text, prosody, and gesture – to create a dense instructional environment. Within this context, viewers must quickly construct an understanding of new vocabulary while simultaneously processing multiple streams of information. The instructional design of TikTok-based language videos, therefore, creates an instructional context distinct from traditional classrooms, in which students are constrained by both the temporal nature of the video format (videos are usually shorter than 60 seconds) and the quantity of multimodal information that they are asked to process during each video (Zeng et al., 2021).

Most research on TikTok and language learning has focused on whether TikTok-based instruction can motivate, engage, and support learners (Alshreef & Khadawardi, 2023; Chuah & Ch'ng, 2023; Waroh et al., 2025). While these studies provide important insights into how learners react to TikTok-based instruction, they rarely address how the instructional design of TikTok-based videos influences student learning.

Although the multimodal nature of TikTok vocabulary content is frequently acknowledged in the literature, few researchers have conducted systematic analyses of how multimodal resources are used in TikTok-based instructional videos to create vocabulary meanings in short-form instructional contexts (Jiang & Hafner, 2025). Prior TikTok language-learning research has concentrated on learner perceptions, engagement, and measured vocabulary gains (Alshreef & Khadawardi, 2023; Chuah & Ch'ng, 2023; Waroh et al., 2025), while the small number of studies that examine the videos themselves have documented which modes are present rather than how those modes are combined and timed (Shanthi et al., 2024). What remains unexamined is the temporal dimension of this design: how densely multiple semiotic cues are clustered and sequenced within the compressed intervals that characterise short-form video. The present study addresses this dimension through the concept of temporal density, defined here as the concentration of multimodal cues, spoken explanation, on-screen text, gesture, and prosodic emphasis, within brief instructional intervals. Temporal density directs analytic attention not to whether modes are present, but to how tightly they converge in time, an aspect of short-form instructional design that existing TikTok research has not yet addressed.

From a multimodality standpoint, meaning is created when users interact with semiotic resources, rather than when using language alone (Kress, 2010; Kress & van Leeuwen, 2001). Therefore, in video-based instruction, linguistic explanation, in combination with visual cues, prosodic variation, embodied gesture, and spatial framing all function together to produce an overall message. While it is possible to identify the presence of multiple modes, it is much more important to understand how those modes are ordered, layered, and synchronised over time to direct user attention and interpretation. As a result, without analysing the orchestration of multimodal elements, multimodality becomes merely descriptive, rather than analytical. A cognitive perspective further emphasises the relevance of multimodal orchestration in learning. Both the Cognitive Load Theory and the Cognitive Theory of Multimedia Learning suggest that the sequencing and density of multimodal cues may shape how instructional input is processed (Mayer, 2005; Sweller, 2011, 2020). Despite the relevance of understanding multimodal orchestration in short-form instructional environments, empirical research examining how such orchestration operates in short-form vocabulary videos is limited.

Due to its semi-informal nature, TikTok occupies a position between formal and informal instructional spaces (Zeng et al., 2021). While learners access vocabulary content outside formal curriculum structures, many of the videos employ pedagogical strategies typical of classroom instruction, such as explicit definitions, pronunciation modelling, example sentences, and repetition. As such, TikTok can be viewed as a semi-informal instructional space that, although intersecting with formal education, retains platform-specific affordances (Zeng et al., 2021; Vizcaíno-Verdú & Abidin, 2023). Understanding how vocabulary meaning is constructed in this environment has implications for digital pedagogy and broader conceptualisations of multimodal instructional design in formal education.

To fill this gap, the current study uses qualitative multimodal discourse analysis to investigate how vocabulary meaning is constructed in TikTok instructional videos. Unlike previous studies that examined learner reactions, this study focused on the instructional texts themselves. By systematically coding the videos, this study examined how multimodal resources interact under short-form temporal constraints.

1.1 Research Questions

The study is guided by the following research questions:

> RQ1: To what extent do TikTok vocabulary videos utilise linguistic, visual, auditory, embodied, and spatial modes in constructing vocabulary meaning?
>
> RQ2: How are these multimodal elements orchestrated over time to construct instructional meaning within short form temporal constraints?

2.0 Literature Review

2.1 Multimodality and Meaning Construction in Digital Instruction

Meaning creation in the context of multimodal theory occurs through the coordinated selection of semiotic resources, rather than solely through the use of language (Jiang & Hafner, 2025; Kress, 2010). Various modes used in digital communication, including linguistic expression, visual representation, embodied movement, auditory variation, and spatial arrangement, provide distinct affordances. Therefore, meaning is constructed not merely through the inclusion of these modes, but through the ways in which they interact with one another over time (Jiang & Hafner, 2025).

Multimodal discourse analysis has been applied to a variety of educational contexts, including classroom interactions, textbooks, and digital media (Jewitt, 2009; Jiang & Hafner, 2025). This research demonstrates that gestures, visual emphasis, and spatial organisation may direct learners' attention and support their comprehension. However, many studies on multimodal discourse analysis focus on cataloguing the presence of modal elements within a text (e.g. identifying that gestures or visual design features occur) rather than analysing how those elements function together across time, particularly within temporally constrained instructional settings such as short-form digital video, where modal sequencing is central to meaning construction. Therefore, in short-form video contexts, the meaning of the text develops rapidly, often within seconds.

TikTok vocabulary videos represent a highly condensed form of multimodal communication. Often, spoken explanations are accompanied by on-screen text, prosody, gestures, and consistent framing. To understand how each of these modal element's function in conjunction with each other, the analyst needs to consider both the interaction among the modal elements and the progression of the modal elements across time (Jiang & Hafner, 2025; Kress, 2010).

2.2 Cognitive Perspectives on Multimodal Instructional Design

Cognitive learning theories provide a way to look at and understand the multimodal (auditory, visual, etc.) instructional environment. Cognitive Load Theory explains that working memory is limited in capacity and that instructional design should therefore manage both intrinsic and extraneous cognitive loads (Sweller, 2020). Poorly coordinated multimodal cues may increase the extraneous cognitive load (Sweller, 2011). Conversely, the strategic coordination of multimodal cues can help learners focus their attention on germane processing while reinforcing related information (Mayer & Fiorella, 2022).

The Cognitive Theory of Multimedia Learning further proposes that meaningful learning occurs through the integration of verbal and nonverbal information across dual processing channels (Mayer, 2005). Signalling, temporal contiguity, and redundancy are examples of principles used to coordinate spoken explanations with visual and auditory cues (Mayer, 2005; Mayer & Fiorella, 2022). These are especially pertinent in short instructional videos, where the learner has to quickly and simultaneously integrate a multitude of inputs in compressed time frames.

While both Cognitive Load Theory and Multimedia Learning principles have been used extensively in semi-informal or informal digital learning environments, these theories have been less frequently applied to short form, platform-based instructional content (Mayer, 2005; Sweller, 2020). Vocabulary videos on TikTok take place in an environment that includes rapid pacing, multimodal layering and high salience (Zeng et al., 2021). Studying how multimodal cues are sequenced and aligned within this type of format will allow for the expansion of cognitive theories of learning into current short-form instructional design. Whereas these theories describe how individual cues such as signalling or temporal contiguity operate, they say less about how densely such cues accumulate within very short intervals, the property this study terms temporal density.

2.3 TikTok as a Semi-Informal Vocabulary Instructional Space

TikTok exists at the intersection of formal educational practices and informal self-directed learning (Vizcaíno-Verdú & Abidin, 2023). Although users engage with educational content voluntarily and outside formal curricula, TikTok vocabulary videos often employ recognisable pedagogical strategies associated with classroom instruction. Therefore, TikTok may be thought of as a semi-informal instructional space where pedagogical intent intersects with platform-specific affordances (Zeng et al., 2021). This positioning is also reflected in studies of teacher-creators on TikTok, who have been found to deploy the platform's affordances to construct educational communities through complex multimodal ensembles that blend pedagogical intent with platform-native performance styles (Vizcaíno-Verdú & Abidin, 2023).

The platform's features also affect the way vocabulary content is presented to users and how it is experienced (Vizcaíno-Verdú & Abidin, 2023). Auto-captions and dynamic text overlays are used to draw attention to key lexical items visually (Vizcaíno-Verdú & Abidin, 2023). Additionally, the "Green Screen" feature allows the creator to provide explanations in front of contextual images or through written examples. Furthermore, audio tools enable creators to emphasise certain parts of an explanation through prosody and repetition. Finally, interactive affordances, such as duets and stitches, allow recontextualisation and repetition across videos, contributing to the high-density and performative nature of TikTok vocabulary instruction (Zeng et al., 2021; Vizcaíno-Verdú & Abidin, 2023). Overall, TikTok vocabulary instruction presents unique aspects that distinguish it from traditional video-based learning environments (Vizcaíno-Verdú & Abidin, 2023).

Although there has been an increased amount of research focused on the use of TikTok as a language learning tool, much of the previous research has focused primarily on learner-centred outcomes. For example, Chuah and Ch'ng (2023) drew on questionnaire-based instruments to capture perceptions of TikTok as a learning tool, while Alshreef and Khadawardi (2023) examined the effectiveness of TikTok-based instruction on EFL vocabulary gains. Waroh et al.'s (2025) systematic review similarly synthesised outcome-oriented findings across multiple studies. While these approaches offer valuable evidence of TikTok's motivational and instructional potential, they do not examine the videos' internal multimodal design. A notable exception is Shanthi et al. (2024), who developed a framework to analyse multimodal elements – including gesture, gaze, typography, and framing – in TikTok vocabulary videos; however, their analysis focuses on identifying the presence of modal features rather than examining how those features are orchestrated across time.

By analysing TikTok vocabulary videos as multimodal instructional texts and foregrounding the concept of orchestration over simple modal presence, this study aims to clarify how meaning is constructed within short-form digital environments and contribute to both multimodality research and cognitive perspectives on multimedia learning.

3.0 Methods

3.1 Research Design

This study employed qualitative multimodal discourse analysis. Multimodal discourse analysis was used because of the unique nature of TikTok vocabulary videos. The meaning of vocabulary is created via the coordination of spoken explanation, visual text, body gesture, prosody, and spatial organisation in these videos. Thus, only a multimodally grounded analysis can adequately capture how instructional

meaning is constructed within short-form digital formats. This study examined instructional design rather than learner outcomes, with the goal of exploring how multimodal resources are organised within time-constrained videos and how they function in the construction of vocabulary meaning (Kress, 2010; Jiang & Hafner, 2025).

3.2 Video Identification and Corpus Development

The dataset comprised ten TikTok videos designed to teach English vocabulary. Ten videos were manually identified and collected from TikTok over a three-month period. The search terms used to locate the dataset were "Learn English", "English Vocabulary", "Vocabulary Lesson", and "English Teacher". To avoid incorporating any potential selection bias into the analysis based on how TikTok uses its algorithm to generate content recommendations for users, personal algorithmic recommendations were disabled during the collection process.

For inclusion in the dataset, each of the ten videos had to satisfy the following four requirements: First, the primary focus of the video was teaching English vocabulary. Second, the video had to be taught in English. Third, the length of the video had to be within the range of typical short-form videos (approximately 30–60 s). Finally, the video had to include visual representations (multimodal elements) of the vocabulary being taught through at least two different semiotic modes.

The study split the video samples equally between the institutional (established organisations) and the creator-driven (individual educator) production processes. The institutional videos were created by English language education organisations with identifiable branding and structured instructional presentation. Most of these institutional videos originated from Britain and other parts of Europe. In contrast, the creator-driven videos were produced independently and were not formally affiliated with an organisation. The creator-driven videos included creators from American, British, and European contexts, which added to the overall variety of prosody, style, and multimedia emphasis throughout the collection of videos. The focus on English-language content from Western institutional and creator contexts was deliberate. The study required a linguistically bounded corpus to enable systematic comparison of multimodal orchestration across videos, and English-language instruction by established institutional and independent creator accounts provided the most consistently documented examples of the vocabulary instruction format under investigation. All videos were publicly accessible at the time of collection and were used solely for non-commercial academic analysis. Table 1 presents an overview of the analysed video corpus.

Table 1

Overview of Analysed TikTok Vocabulary Videos

Video ID	Producer Type	Region	Duration	Vocabulary Focus
V1	Institutional	United Kingdom	42 sec	Everyday nouns
V2	Institutional	United Kingdom	38 sec	Irregular plurals
V3	Institutional	Europe	45 sec	Descriptive adjectives
V4	Institutional	Europe	51 sec	Informal expressions
V5	Institutional	United Kingdom	34 sec	Common collocations
V6	Creator-Driven	Europe	56 sec	Phrasal verbs
V7	Creator-Driven	Europe	49 sec	Everyday expressions
V8	Creator-Driven	Europe	37 sec	Synonyms
V9	Creator-Driven	Europe	58 sec	Food vocabulary
V10	Creator-Driven	Europe	43 sec	Health expressions

Theoretical saturation was used to determine the number of videos to be coded for this study. Through iterative coding (i.e., repeated analysis) by the eighth video, the different types of multimodal data categories began to stabilise and were fully stabilised by the tenth video, with no new categories or orchestration patterns emerging. Coding of subsequent candidate videos revealed that the categories established through the first ten videos were consistently observed but never expanded; therefore, data collection was concluded at that time.

The decision to analyse ten videos also reflects the analytic demands of multimodal discourse analysis. Because each video was coded at the level of individual timestamped segments across five semiotic modes, the corpus generated 218 discrete coded quotations, a level of fine-grained analysis that favours a small, purposively selected corpus over a large one (Jewitt, 2009). Small corpora of this kind are common in multimodal discourse analysis, where the aim is analytic depth rather than statistical generalisation. The corpus was therefore not designed as a representative sample but as a purposively balanced set intended to support structured comparison. The five institutional and five creator-driven videos were selected to hold the instructional genre constant while varying the production context, so that any differences in multimodal orchestration could be examined against a common analytic frame. The institutional and creator-driven contrast is accordingly interpreted as an analytic distinction internal to this corpus rather than as a generalisable claim about the two categories.

3.3 Multimodal Coding Framework and Analytical Procedures

A multimodal coding framework was developed to analyse the ways in which vocabulary meaning is created using five semiotic modes (linguistic, visual, auditory, embodied, spatial) as defined within multimodality theory and cognitively based research on multimedia learning. Multimodality theory emphasises that meaning construction occurs when multiple modes interact with one another. Similarly, multimedia learning research suggests that meaning is constructed through multimodal interaction.

The initial coding framework for this study was developed deductively from these theories, and subsequent refinements were made inductively after viewing the video corpus several times. Refinements included defining the boundaries of each code, combining overlapping codes, and developing codes that capture an analytical understanding of each mode's instructional function as opposed to their surface characteristics. Table 2 summarises the multimodal framework used in this study.

Table 2

Multimodal Coding Framework

Mode	Operational Definition	Example Indicators
Linguistic	Verbal and written lexical content used to introduce or explain vocabulary items	Explicit word naming, repetition, example sentences, definition statements
Visual	Graphical or textual elements displayed on screen to support lexical recognition or emphasis	On screen text, subtitles, colour highlighting, animated overlays
Auditory	Acoustic features shaping salience and instructional pacing	Prosodic stress, intonation shifts, rhythmic pacing, emphasis patterns
Embodied	Physical movement or facial expression supporting semantic explanation or emphasis	Beat gestures, iconic gestures, pointing, facial emphasis
Spatial	Organisation of visual space and framing within the video	Central framing, text placement, gesture space allocation, visual hierarchy

Videos were broken down into separate instructional sections to analyse the variation in modal density and orchestration at different instructional moments versus treating the videos as units of instruction.

Modal occurrences and co-occurrences were coded using ATLAS.ti for timestamped analysis of modal convergence within brief time frames. The use of this software allowed researchers to systematically map out areas (and times) that multiple modes co-occurred within short timeframes. Modal coding was done by one researcher that had been trained in multimodal analysis. It is acknowledged that single-coder analysis without intercoder reliability testing represents a methodological limitation of this study. Within interpretive qualitative research, however, trustworthiness is established through alternative criteria rather than statistical agreement measures. Accordingly, the researcher engaged in continued memoing of their analytic work and iterative refinement of the code definitions (Lincoln & Guba, 1985). To strengthen analytical trustworthiness, the primary coder engaged in peer consultation with two applied linguists (Dr Rachel Tan Sing Ee and Dr Kumutha Raman), both of whom hold expertise in multimodal discourse analysis. Each consultant independently reviewed a purposively selected subset of coded video segments, specifically those identified as containing complex or ambiguous co-occurrence patterns and discussed their interpretations with the primary coder. The peer-consultation process followed a consistent procedure. For each flagged segment, the consultant and the primary coder first coded the segment independently, then compared their assignments and discussed any divergence until a shared interpretation was reached. Where a divergence reflected an unclear code boundary rather than a simple oversight, the relevant code definition was revised, and the affected segments were recoded under the revised definition. For example, one recurring point of divergence concerned centrally held hand movements that could be read either as an embodied cue (a beat or iconic gesture) or as a spatial cue (use of a central gesture space); this was resolved by refining the definitions so that the embodied codes captured the movement itself while the spatial codes captured its framing within the visual field, allowing both features to be recorded without double-counting the same cue. A further point of divergence involved separating prosodic stress used to segment speech from stress used to mark a new lexical item, which led to sharper definitions within the auditory mode. These revisions were then applied across the corpus, and coding decisions were documented throughout so that the development of the framework could be traced. This consultative process, while not constituting formal intercoder reliability testing, provided an external check on analytical consistency and contributed to the credibility of the coding outcomes (Lincoln & Guba, 1985). The researcher also kept documentation of coding decisions so that their analytic development could be clearly demonstrated.

3.4 Operationalising Temporal Density

Temporal density is conceptualised as the concentration and rapid sequencing of multimodal cues within limited instructional timeframes. It refers to the degree to which spoken explanations, visual emphasis, gestures, and auditory cues converge within short intervals. This construct builds on cognitive accounts of multimedia learning, particularly temporal contiguity and the management of cognitive load, while extending them to the rate at which cues accumulate (Mayer, 2005; Sweller, 2011).

Temporal density was operationalised through timestamped co-occurrence analysis in ATLAS.ti software. The segments were examined to identify moments when multiple semiotic modes overlapped or were introduced in rapid succession. Attention was paid to the vocabulary introduction and reinforcement phases, where cue clustering was most pronounced. This operationalisation enabled a systematic examination of how multimodal compression may be interpreted in relation to cognitive processing demands (Jewitt, 2009).

It is important to differentiate temporal density from other constructs in the multimedia learning literature. Pacing refers to speech rate and delivery speed – a unimodal auditory feature – and does not account for the convergence of multiple modes within the same interval. Signalling describes design principles used to guide the learner's focus on specific elements through cues such as arrows, highlighting, or emphasis (Mayer & Fiorella, 2022); however, it does not consider how densely or rapidly those cues accumulate over time. Similarly, redundancy refers to the intentional presentation of equivalent information using more than one mode to support reinforcement (Mayer, 2005). However, redundancy does not account for the temporal compression within which such reinforcement occurs. Temporal density refers to the concentration of multiple semiotic cues within a narrow instructional

interval. Pacing concerns delivery speed, whereas signalling and redundancy describe design functions. In contrast, temporal density captures the extent to which such cues accumulate and converge over time. As a result, temporal density is a product of the relationship between pacing, signalling, and redundancy, while operating under temporal constraints; it is not synonymous with any one of them (Mayer, 2005; Mayer & Fiorella, 2022).

4.0 Results

The results indicate that the meaning of vocabulary in the corpus was represented through recurring combinations of linguistic, visual, auditory, embodied, and spatial resources. Across the ten videos, linguistic explanation formed the instructional base, while lexical salience was commonly reinforced through visual, auditory, embodied and spatial cues.

4.1 Modal Distribution Across the Corpus

The auditory and linguistic modes were the most frequently observed across the corpus. Each video included recurring instructional patterns, such as spoken word introductions, repetition sequences, contextual examples, and prosodic emphasis on key words. This pattern is evident in both institutional and creator-driven videos, where lexical items are commonly introduced through spoken prompts, isolated target words, and prosodic emphasis.

Visual, embodied, and spatial modes were used as supporting materials to emphasise specific lexical items and clarify aspects of meaning. The visual text overlays were always synchronised with the lexical introduction, which made use of subtitles or highlighted captions at the exact time the target word was pronounced, thereby providing additional support for the recognition of the orthography of the new word. Auditory supports such as stress, pacing, and intonation were commonly observed during key lexical presentations. In creator-driven videos, brief sound effects sometimes coincide with the appearance of new visual elements, functioning as auditory markers of lexical transition rather than content carriers.

The embodied component of the videos included beat and iconic gestures which were primarily shown while introducing and reinforcing new vocabulary. The spatial frame of reference remained relatively constant and centrally located throughout most of the videos, with minimal distractions from the visuals. In some creator-driven videos, animated visual props and labelled graphics appeared alongside spoken vocabulary during introductory segments, with on-screen labels synchronising with the corresponding spoken forms. Table 3 presents the frequency distribution of all active codes across the corpus, organised by semiotic mode. Frequency refers to the number of coded quotations (discrete timestamped segments) in which each feature was identified across all ten videos.

Table 3

Code Frequency Distribution by Semiotic Mode Across the Ten-Video Corpus (N = 218 coded quotations)

Semiotic Mode	Code	Frequency	Institutional (V1-V5)	Creator-Driven (V6-V10
Auditory	Slow-paced for comprehension	9	2	7
	Stress on key syllables	8	4	4
	Clear articulation	6	3	3
	Repetition instruction	4	4	0
	Accompanying background clicks	4	2	2
	Calm, steady tone	4	0	4
	Strategic pauses	4	1	3
Embodied	Beat gestures	9	4	5
	Reinforcing gesture	6	4	2

Semiotic Mode	Code	Frequency	Institutional (V1-V5)	Creator-Driven (V6-V10
Visual	Gesture marking introduction	5	3	2
	Iconic gestures	5	2	3
	Rhythmic body posture	5	1	4
	Open posture	5	2	3
	Direct gaze	8	2	6
	Encouraging smile	7	4	3
	Friendly neutral expression	6	3	3
	Props	5	0	5
	Subtitles	4	0	4
Spatial	Central gesture space	6	3	3
	Stable frame	5	4	1
	Consistent orientation	5	3	2
Linguistic	Clear word introduction	4	1	3
	Conversational	4	2	2
	Final repetition	3	3	0
	Key examples	6	3	3
	Total coded quotations	**218**	**98**	**120**

Several patterns in Table 3 are notable. First, slow-paced delivery and beat gestures were the two most frequently coded features overall (each n = 9), indicating that pacing control and embodied rhythm are central to foregrounding vocabulary across both video types. Second, creator-driven videos accounted for a substantially higher proportion of deliberate slow-pacing at key instructional moments (n = 7 vs. n = 2), direct gaze (n = 6 vs. n = 2), props (n = 5 vs. n = 0), and subtitles (n = 4 vs. n = 0), reflecting a more expressive multimodal design orientation. In contrast, institutional videos showed a stronger concentration of stable framing (n = 4 vs. n = 1), reinforcing gestures (n = 4 vs. n = 2), and final repetition (n = 3 vs. n = 0), suggesting a preference for structural clarity over platform-native expressiveness. Third, Video 4 recorded the lowest coded quotation count in the corpus (n = 5, compared to a corpus mean of 21.8), reflecting the comparatively minimal multimodal design of the institutional clip.

4.2 Multimodal Co-Occurrence Patterns

Co-occurrence analysis using ATLAS.ti revealed that the patterns of modal overlap were consistent and widespread across the corpus. Specifically, the most consistent pairing was linguistic–visual; spoken lexical naming, for example, was usually shown at the same time as a textual display on-screen when a new word was being introduced and there was an association to be made. This paired mode was observed almost every time a lexical item was introduced and functioned to support the recognition of word forms and reduce orthographic ambiguity.

The next most common pairing was linguistic–auditory pairing. Prosody, including rhythm, pacing, and pauses, for example, often coincides with the division of spoken language into individual lexical items. Prosody in these segments foregrounded phonological features and provided a model for pronunciation. Linguistic explanations and embodied gestures also frequently co-occurred. In this case, beat gestures often marked syllabic boundaries, while iconic gestures appeared to externalise the semantic content of the lexical item. The degree to which these two types of gestures overlapped with linguistic explanations was especially evident in the early stages of the lexical introduction process.

There were also analytically notable clusters of modes, where three modes occurred together within relatively short periods. For example, spoken naming of the lexical item, highlighting the textual representation of the item, and accompanying gestures would often occur near one another to create high levels of multimodal density. These clusters of modes, however, were typically located around the

point of first introduction of new vocabulary and the subsequent reinforcement of that vocabulary and not dispersed randomly throughout the videos.

Appendix A presents the ten strongest co-occurrence pairs identified through the ATLAS.ti co-occurrence analysis, ranked by the number of segments in which both codes were active simultaneously.

The co-occurrence data in Table A1 reveal a consistent pattern of cross-modal pairing centred on embodied and spatial resources. Beat gestures co-occurred most strongly with direct gaze, stable framing, and central gesture space, indicating that the embodied activity was not performed in isolation but was systematically anchored within a stable visual field directed at the learner. This tripartite convergence of embodied movement, spatial predictability, and sustained eye contact represents the most recurrent multimodal configuration in this corpus. This configuration may function as a form of spatial predictability: by keeping the face and hands in a consistent central field, creators appear to limit the area of the frame that the learner must monitor, which could in turn make it easier to process rapidly changing text overlays without re-scanning the whole frame. Because no learner data were collected, this reading is offered as a design-level inference rather than a demonstrated effect on attention or processing.

The second notable pattern concerns the pairing of direct gaze with slow-paced delivery (co-occurrence = 5), suggesting that moments of deliberate auditory deceleration were consistently accompanied by intensified visual engagement, a design pattern that concentrates visual and auditory emphasis at points where phonological and semantic information is made most explicit. The co-occurrence of clear articulation with direct gaze (co-occurrence = 3) reinforces this interpretation, as pronunciation modelling was characteristically delivered with sustained eye contact rather than with redirected attention. Together, these patterns indicate that co-occurrence in this corpus was neither random nor incidental but reflected systematic cross-modal coordination at instructionally significant moments in the lesson.

To assess whether co-occurrence patterns reflected systematic cross-modal coordination rather than isolated pairings, co-occurrence counts were aggregated at the semiotic mode level. Table 4 presents the total co-occurrence instances for each pair of modes across the corpus.

Table 4

Cross-Modal Co-occurrence Summary: Total Instances of Code Co-occurrence Between Semiotic Modes

Mode	Auditory	Embodied	Visual	Spatial	Linguistic
Auditory	-	38	29	18	24
Embodied	38	-	41	35	22
Visual	29	41	-	19	20
Spatial	18	35	19	-	14
Linguistic	24	22	20	14	-

Table 4 reveals that the strongest cross-modal relationships in the corpus were between the embodied and visual modes (n = 41) and between the embodied and auditory modes (n = 38). This pattern is consistent with the co-occurrence pair analysis in Table A1 where beat gestures co-occurred most strongly with direct gaze and slow-paced delivery. The embodied mode therefore functions not as an isolated channel of meaning but as an integrative hub through which visual and auditory resources are coordinated. The spatial–embodied relationship (n = 35) reflects the stable framing that consistently contained embodied activity, as discussed in Section 4.2. By contrast, the linguistic mode showed the lowest co-occurrence totals with spatial (n = 14) and embodied features (n = 22), consistent with the argument in Section 5.1 that the linguistic mode served as the instructional anchor but was not the

primary vehicle for multimodal orchestration. The orchestration of meaning in this corpus therefore operated most intensively across three modes – embodied, visual, and auditory – rather than radiating evenly from language outwards.

Table 5

Dominant Multimodal Co-Occurrence Patterns

Modal Pair or Cluster	Relative Strength	Instructional Function
Linguistic + Visual	Highest	Orthographic reinforcement and form recognition
Linguistic + Auditory	Strong	Prosodic salience and segmentation
Linguistic + Embodied	Frequent	Semantic amplification through gesture
Visual + Embodied	Strong	Deictic attention guidance
Linguistic + Visual + Embodied	Recurrent cluster	High salience lexical introduction

These results indicate that TikTok vocabulary videos employ structured multimodal convergence rather than the random layering of semiotic resources.

4.3 Integrated Multimodal Themes

Beyond modal pairings, the analysis identified broader integrative themes that characterised the instructional design across the corpus. These themes capture how multiple semiotic resources functioned collectively to create meaning.

Table 6

Multimodal Integration Themes in TikTok Vocabulary Lessons

Integrated Theme	Modes Involved	Functional Description
Attention Management	Auditory, Visual, Spatial	Stress, text overlay, and framing guide learner focus
Prosodic Marking	Auditory, Linguistic	Emphatic stress and intonation enhance lexical salience
Gesture–Speech Synchrony	Embodied, Auditory, Spatial	Coordinated movement aligns with syllable segmentation
Iconic and Deictic Meaning Making	Embodied, Visual, Linguistic	Physical representation supports semantic mapping
Redundant Multimodal Marking	Linguistic, Visual, Auditory	Multiple cues converge to reinforce a single lexical item

A primary pattern observed across the corpus was the systematic deployment of convergent cues to facilitate the management of attention. Prosodic emphasis was typically used in conjunction with both highlighting the text and stabilising the frame. The coordination of these three modalities (visual, auditory, and spatial) functions to direct attention toward critical lexical items while minimising competing visual information.

Another prominent pattern was gesture–speech synchrony. The beat gestures were often synchronised with stressed syllables, while the introductory gestures signalled when a new word would be introduced.

Iconic gestures provided an additional way for individuals to represent meaning visually; specifically, they assisted in providing a physical representation of the meanings of concrete vocabulary items. These embodied cues appeared to contribute to semantic representation by providing a physical analogue of lexical meaning.

Another pattern involved redundant multimodal marking, where lexical items were reinforced simultaneously through spoken naming, textual display, and prosodic emphasis. These cues did not necessarily introduce new semantic information, but they increased the salience of the target vocabulary across multiple channels.

4.4 Temporal Density and Instructional Compression

Temporal density emerged as an important characteristic of TikTok vocabulary lessons. High-density sections contained high levels of temporal convergence in oral explanation, textual highlighting, gesture, and auditory emphasis. For example, in Video 6, deliberate slow pacing at the opening (00:00:00.000–00:00:02.796) was followed by vocabulary overlays and explanations appearing in close succession (00:00:06.843–00:00:14.349), compressing instructional phases with limited consolidation time between them. Specifically, in Video 6, a deictic gesture (pointing) at 00:08.2 appears to function as a pre-cue, positioned to direct attention toward the top-right quadrant of the frame roughly 150 ms before the word 'Ecstatic' appears as a text overlay. This micro-alignment may reflect a deliberate staggering of cues that could reduce visual search time and thereby help manage the high temporal density of the segment, although any such effect on viewer processing remains an inference rather than a measured outcome. By contrast, Video 9 incorporated a brief pause between explanation and example (00:00:32.172–00:00:38.388), creating a brief interval between explanation and example that reduced multimodal clustering. This contrast illustrates that temporal density is not uniform across the corpus but varies according to design choices made by individual creators.

Creator-driven videos exhibited greater multimodal clustering and faster pace than institutional videos. They also exhibited more dynamic gestures and greater variations in pitch and volume. In contrast, institutional videos tended to maintain a slower pace and clearer separation between the explanation and example phases.

5.0 Discussion

This study aimed to investigate how vocabulary meaning is produced in TikTok instructional video content through multimodal orchestration. The results suggest that multimodal presence alone does not fully account for meaning making in instruction. Rather, meaning construction in short-form digital contexts appears to depend on the coordinated patterning of modes, particularly at moments of lexical introduction and reinforcement.

5.1 Extending Multimodality Beyond Modal Presence

Across the corpus, the linguistic mode formed a consistent instructional base; however, lexical prominence was most often achieved through the simultaneous convergence of visual text, prosody, and/or gesture. Although these moments of convergence occurred simultaneously throughout the video, they did not occur randomly throughout the video. Instead, the majority of the moments of convergence occurred at the same instructional points; therefore, it appears that TikTok vocabulary videos use modal design strategically through the selective activation of multimodal features and not by continuously layering them. This selection in terms of function also reflects the practical division of functions across modes. While the linguistic mode was consistently used as an anchor for introducing new vocabulary into the lesson, it was the combination of auditory, visual, embodied, and spatial resources at instructional times (especially during vocabulary introduction and final reinforcement) which most strongly amplified lexical salience. It was possible to present the word form and general meaning using only the linguistic mode. However, it was the coordinated occurrence of prosody, on-screen text, and embodied representation of words at the exact same time and location as these strategic moments in instruction that clearly differentiated pedagogically emphasised items from background explanatory information. This pattern supports the argument that multimodal orchestration is not merely additive but specifically focused on strategically selected points in time.

Therefore, the findings contribute to multimodal theory by emphasising that temporal ordering is key to creating meaning in short-form content. The modal interactions in these videos do not exist statically; instead, they are dynamically constructed over time.

5.2 Temporal Density as an Instructional Variable

Temporal density emerged as one of the most theoretically significant analytical patterns in this study. Although well-known instructional design principles such as pacing, signalling, and redundancy have been established in the cognitive learning literature, the analysis suggests that these principles often co-occur in short-form instructional contexts. In these videos, they often occurred together within very short instructional time frames, and their co-occurrence within narrow instructional windows produced temporal density.

Two contrasting temporal density patterns were observed across the corpus. Video 6 was characterised by slow-paced introductory segments (00:00:00.000 – 00:00:02.796) followed by a rapid series of vocabulary overlays and spoken explanations (00:00:06.843–00:00:14.349). Overlays, spoken naming, visual texts, prosodic emphasis, and gestures co-occurred within a tightly compressed temporal window.

Video 9 represents an alternative approach. The creator incorporated a strategic pause between the explanation and example phases (00:00:32.172 – 00:00:38.388), thereby creating a brief interval that may have afforded greater time for the integration of multimodal content. Similar patterns of moderate density were evident in Videos 8 and 9. These videos displayed a more regular pace and lower tonal variation, resulting in less dense multimodal clustering. However, this pattern reduced the dynamic emphasis. Taken together, these contrasting examples illustrate that temporal density is a design choice made by individual creators working in the short-form context and is not inherent to the short-form format per se.

These patterns have theoretical significance for how differences in temporal density and instructional rhythm may affect cognitive processing; however, due to their speculative nature, they will be discussed at a theoretical level rather than empirically demonstrated. Sweller's (2020) Cognitive Load Theory posits that learners can process a limited amount of information at any given time and that managing both intrinsic and extraneous cognitive loads is key to creating effective instruction. Mayer's (2005) work on temporal contiguity suggests that the simultaneous presentation of verbal and visual information facilitates dual-channel processing. The findings of this study suggest that when multimodal cues are well-coordinated, the benefits of temporal contiguity may depend not only on simultaneity but also on how densely multimodal cues are clustered within a short interval. Although the data identify instances of cue clustering, the specific cognitive impact remains speculative. Instances in which multiple cues appeared simultaneously without a clear focal point may theoretically lead to attentional competition; however, without learner data, these remain tentative design observations rather than confirmed cognitive outcomes.

Although the cognitive interpretations provided above represent plausible inferences, they remain theoretical. Since this study was focused on designing instruction rather than learner cognition during instruction, no comprehension or processing measures were collected as part of this study. Experimental studies involving student performance data provide evidence of whether the temporal density patterns identified in this study produce the cognitive effects described above. Therefore, the contribution of this study is that temporal density is an identifiable and analytically distinguishable feature of short-form vocabulary instruction and that temporal density varies systematically across videos and video types. Additionally, variations in temporal density reflect deliberate design decisions made by instructors rather than being solely a matter of style. Establishing this foundation is essential for further enquiry into the cognitive outcomes associated with temporal density.

5.3 TikTok as a Semi-Informal Instructional Space

An additional contribution of this analysis is toward the understanding of TikTok as a semi-informal instructional space (Zeng et al., 2021). While the videos exist outside formal curriculum structures, they use recognisable pedagogical strategies (such as explicit definition, repetition, segmentation, and example modelling) (Zeng et al., 2021). Simultaneously, platform affordances (such as dynamic text

overlays, gestures, and pace) create a design that does not typically appear in a traditional classroom instructional setting.

This distinction between institutional and creator-driven videos also demonstrates the hybrid nature of this space. Institutional videos had more measured pacing and segmented video content, while creator-driven videos were denser in terms of multimodality and contained more elements of performance. This contrast is not merely stylistic in nature. Institutional videos (Videos 1-5) exhibited single-mode dominance or limited multimodal convergence, prioritising clarity and structural stability over expressive richness. Creator-driven videos (Videos 6-10), by contrast, demonstrated dense alignment across all five modes simultaneously, particularly at moments of vocabulary introduction, reflecting creators' familiarity with TikTok's affordances and the platform's expectations of immediacy, engagement, and visual salience. Gesture-speech synchrony, dynamic visual emphasis, and deliberate use of central frame space were consistently more prominent in creator-driven content than in brand-driven content. In short, creator-driven instructors do not simply teach on TikTok; they design their content with the platform's logic in mind more consistently than institutional producers in this corpus.

This empirical contrast maps directly onto the conceptualisation of TikTok as a hybrid instructional space. The divide between institutional and creator-driven content visible in this corpus is not simply a difference in production quality or pedagogical intent. This reflects two distinct orientations toward the platform: one that imports formal instructional conventions into a digital space and one that builds instruction from the platform's own affordances outwards (Zeng et al., 2021). Rather than positioning TikTok as either informal or instructional, the present findings suggest that it functions as an interface between these logics, where each content type appears to offer different strengths: institutional content provides cognitive manageability and predictability, while creator-driven content provides the multimodal richness and engagement that the platform's audience expects.

5.4 Implications for Multimodal Instructional Design

The research shows that effective multimodal design in short-form vocabulary videos appears to depend more on the coordination of modes than on the sheer variety of modes employed. Lexical naming in synchrony with text may support orthographic recognition. Phonetic prominence reinforced through gestures appears to enhance lexical salience. Consistent spatial organisation has the potential to reduce visual interference and competition for learners' attention.

In addition, these patterns illustrate that it is the orchestration of multiple modes, especially when a new word is introduced, that appears central to effective instructional design in video-based vocabulary contexts. Therefore, designers of short-form video-based vocabulary instruction may be more successful by focusing on aligning cues temporally and grouping them rather than increasing the visual complexity.

5.5 Limitations and Future Directions

This study analysed ten videos selected to achieve theoretical saturation. Saturation is useful for developing analysis depth; however, the sample does not represent the full range of possible TikTok vocabulary content. The videos were limited to English-language instructional videos found within British, American, and European contexts. Researchers may wish to investigate a wider variety of languages and cultures.

This study analysed instructional texts rather than learner outcomes, and creator intention was inferred from recurring design patterns rather than directly verified. Therefore, temporal density should be understood at this stage as a proposed analytic construct requiring further validation in learner-based research. While several multimodal orchestration patterns were observed, the theoretical inference of the cognitive effects of high temporal density was based on theoretical models and has not been empirically tested through experimental designs. Therefore, researchers wishing to further explore how multimodal compression affects the processing and retention of vocabulary may choose to incorporate student comprehension data into their research.

6.0 Conclusion

This study has shown how vocabulary meaning is constructed in TikTok educational videos through multimodal orchestration. Through the use of a structured multimodal framework and co-occurrence analysis of ten videos, this study moves from identifying the presence of different modes, namely linguistic, visual, auditory, embodied, and spatial, to examining how those modes support one another at particular instructional moments.

The data show that effective vocabulary instruction in short-form digital media appears to involve the coordinated use of the available modes. The most common mode was a combination of the linguistic and visual modes; however, the use of prosody and gestures appeared to enhance the salience of individual words. Furthermore, the data showed that when all three modes (linguistic, visual, and auditory) were combined at the same point in time, this was often associated with the introduction of new vocabulary and therefore highlighted the importance of temporal density as an instructional factor. Multimodal cues, including colour, typography, music, and gestures, were found to operate as part of an integrated signalling system.

This study contributes to multimodality research by highlighting the importance of temporal sequencing and density as key factors in meaning-making in short-form environments. Additionally, this study builds on existing cognitive theories of multimedia learning by suggesting that cue concentration and cue alignment should be considered when assessing cue usage in addition to channel usage. The research also demonstrates that TikTok functions as a semi-informal instructional space in which content design is shaped by both the pedagogical intentions of the creator and the affordances of the platform.

Although the research is limited by its small sample of ten videos, it provides a clear and theoretically grounded description of how vocabulary is multimodally presented in contemporary digital media. As the production and use of short-form instructional videos continue to grow, so too does the need to understand the underlying principles of multimodal orchestration so that researchers and educators can better understand and evaluate short-form instructional design.

Acknowledgement

The authors wish to thank Dr. Rachel Tan Sing Ee and Dr. Kumutha Raman for their expert consultation during the multimodal coding process. We also gratefully acknowledge the support of INTI International University, Universiti Pertahanan Nasional Malaysia, and Universiti Teknologi MARA (UiTM) during this research.

Conflict of Interest

The authors declare that no competing interests exist.

Author Contribution Statement

JX: Conceptualisation, Data Curation, Methodology, Formal Analysis, Validation, Writing – Original Draft Preparation. NZS: Supervision, Project Administration, Writing - Review and Editing. MHK: Supervision, Writing - Review and Editing. JB: Writing - Review and Editing. AS: Validation, Writing - Review and Editing.

Funding

This research received no external funding.

Ethics Statement

In this study, no Institutional Review Board (IRB) review was needed to analyse public TikTok videos because no participants were recruited to interact with the researcher or collect any personal information from the public. The content analysed was collected from the internet and was publicly viewable at that time; thus, it was analysed for academic research purposes only. To be considerate of the academic use of publicly available content, the regional office for TikTok in Singapore was notified of this study.

Data Access Statement:

Research data supporting this publication are available upon request to the corresponding author.

Author Biography

Jane Xavierine is a Lecturer and doctoral researcher in applied linguistics, specialising in multimodal discourse analysis, digital language pedagogy, and vocabulary instruction in short-form media environments.

Assoc. Prof. Dr. Norshima binti Zainal Shah is a Malaysian academic at UPNM specialising in English language education, critical thinking, intercultural communication, and employability skills, with extensive teaching, research, and leadership experience.

Dr. Mohd Hasrul bin Kamarulzaman is a Senior Lecturer at Universiti Pertahanan Nasional Malaysia and currently serves as the Director of the APEL Centre. His supervision spans undergraduate and postgraduate research, with interests in language, culture, communication, media, and education.

Assoc. Prof. Dr Alice Shanthi is from the Academy of Language Studies. Her area of expertise is teaching and learning English, and her interest lies in computer-mediated discourse, especially in the academic field. She has published papers in indexed and non-indexed journals and has presented her research at many national and international conferences. She is also the author of several books used in the UiTM system.

Jonathan Bryce is a Lecturer in English proficiency specialising in IELTS and learning through video media. He is also a doctoral researcher studying how foreigners learn and communicate in Bahasa Melayu, particularly with regard to mobile learning applications.

References

Alshreef, N. R., & Khadawardi, H. A. (2023). Using TikTok as a tool for English vocabulary learning in the EFL context. *English Language Teaching, 16*(10), 125. https://doi.org/10.5539/elt.v16n10p125

Chuah, K. M., & Ch'ng, L. C. (2023). The usefulness of TikTok voice-over challenges as ESL speaking activities: A case study on Malaysian undergraduates. *Electronic Journal of Foreign Language Teaching, 20*(1), 37-49. https://doi.org/10.56040/kmlc2013

Jewitt, C. (Ed.). (2009). The Routledge Handbook of Multimodal Analysis. Routledge.

Jiang, L., & Hafner, C. (2025). Digital multimodal composing in L2 classrooms: A research agenda. *Language Teaching, 58*(4), 528–546. doi:10.1017/S0261444824000107

Kress, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. Routledge

Kress, G., & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Arnold

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.004

Mayer, R. E., & Fiorella, L. (2022). Principles for managing essential processing in multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781108894333.025

Shanthi, A., Jumaat, N. A. F., Suppiah, P. C., Johar, E. M., & Arumugam, N. (2024). From trending to teaching: A framework to analyse TikTok videos for vocabulary instruction. *International Journal of Social Science Research*, 12(2), 136–150. https://doi.org/10.5296/ijssr.v12i2.21904

Sweller, J. (2011). *Cognitive load theory*. In J. Sweller, J. J. G. van Merriënboer, & F. E. Paas (Eds.), *Cognitive load theory* (pp. 37-76). Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development, 68*, 1–16. https://doi.org/10.1007/s11423-019-09701-3

Vizcaíno-Verdú, A., & Abidin, C. (2023). TeachTok: Teachers of TikTok, micro-celebrification, and fun learning communities. *Teaching and Teacher Education*, 123. https://doi.org/10.1016/j.tate.2022.103978

Waroh, S., Pusfitasari, L., & Tarihoran, N. (2025). A systematic review of TikTok for English language teaching. *International Journal of Society Reviews, 3*(5). https://injoqast.net/index.php/INJOSER/article/view/190

Zeng, J., Abidin, C., & Schäfer, M. T. (2021). Research perspectives on TikTok and its legacy apps: Introduction. *International Journal of Communication, 15*, 3161–3172. https://ijoc.org/index.php/ijoc/article/view/14539/3494

APPENDIX A

Strongest Multimodal Co-occurrence Pairs Identified Across the Corpus

Rank	Code A	Code B	Co-occurrence Count	Functional Interpretation
1	Beat gestures	Direct gaze	5	Embodied–visual convergence; combined attentional anchoring
2	Direct gaze	Slow-paced for comprehension	5	Visual–auditory alignment during delivery of key lexical content
3	Beat gestures	Central gesture space	4	Embodied–spatial pairing; gesture kept within stable visual field
4	Beat gestures	Encouraging smile	4	Embodied–visual convergence during energetic instructional moments
5	Beat gestures	Stable frame	4	Embodied–spatial pairing; movement contained within consistent framing
6	Consistent orientation	Gesture marking introduction	4	Spatial–embodied alignment at moments of new vocabulary introduction
7	Direct gaze	Encouraging smile	4	Sustained visual engagement combined with affective signalling
8	Reinforcing gesture	Stable frame	4	Spatial stability maintained during embodied emphasis of lexical items
9	Repetition instruction	Stable frame	4	Linguistic repetition supported by spatial consistency
10	Clear articulation	Direct gaze	3	Auditory–visual pairing during pronunciation modelling sequences

Note. Co-occurrence count refers to the number of coded segments in which both features were simultaneously active, as computed by ATLAS.ti. Pairs involving near-synonymous codes (Iconic gestures / Iconic gestures and speech) are treated as a single construct and excluded.