MULTIMODAL MECHANISMS TO ENGAGE VIEWERS IN APPRECIATING VIETNAMESE STREET FOOD IN A FOOD REVIEW VIDEO

Nguyen Hong Lien⌧**; Duong Thu Ha**

Faculty of English, Hanoi National University of Education, Hanoi, Vietnam

⌧ nguyenhonglien@hnue.edu.vn

https://doi.org/10.63506/jilc.0903.357

(Received: 22/09/2025; Revised: 12/11/2025; Accepted: 29/12/2025)

Abstract: This study investigates how verbal, visual and paralinguistic modes orchestrate to engage viewers in appreciating Vietnamese street food in the most-viewed food review video on YouTube about this cuisine. Using a multimodal discourse analytical framework informed by Systemic Functional Linguistics, the analysis reveals that meaning emerges not from any single mode but from flexible multimodal orchestration, operating through both intersemiosis (the integration of meaning across modes) and intrasemiosis (the elaboration of meaning within modes). Five key mechanisms that drive viewer engagement include creating interpersonal closeness, claiming expert authority, explaining food values, using familiar comparisons and constructing sequential narrative progression. These strategies collectively meet the genre demands of food review videos, which balance personal connection, credibility, educational clarity, accessibility and narrative flow. The study contributes theoretically by demonstrating the applicability of SFL-based multimodal analysis to video data, and practically by offering insights for content creators on how deliberate multimodal coordination enhances audience engagement and cultural appreciation.

Keywords: Multimodal orchestration; verbal; visual; paralanguage; Vietnamese street food; viewer engagement

CƠ CHẾ ĐA PHƯƠNG THỨC TRONG VIỆC THU HÚT NGƯỜI XEM THƯỞNG THỨC ẨM THỰC ĐƯỜNG PHỐ VIỆT NAM TRONG MỘT VIDEO ĐÁNH GIÁ ẨM THỰC

Tóm tắt: Nghiên cứu này điều tra cách các phương thức ngôn từ, hình ảnh và cận ngôn từ phối hợp để thu hút người xem thưởng thức ẩm thực đường phố Việt Nam trong video đánh giá ẩm thực có lượt xem cao nhất trên YouTube về chủ đề này. Sử dụng khung phân tích diễn ngôn đa phương thức (MDA) dựa trên Ngôn ngữ học Chức năng Hệ thống (SFL), kết quả của nghiên cứu cho thấy ý nghĩa không chỉ phát sinh từ một phương thức đơn lẻ mà là từ sự phối hợp đa phương thức linh hoạt thông qua liên phương thức (intersemiosis – tích hợp ý nghĩa giữa các phương thức) và nội phương thức (intrasemiosis – triển khai ý nghĩa trong cùng một phương thức). Năm cơ chế chính thúc đẩy sự tương tác của người xem bao gồm: thiết lập sự gần gũi liên nhân xưng, khẳng định quyền uy chuyên gia, giải thích giá trị ẩm thực, sử dụng các phép so sánh quen thuộc và xây dựng diễn tiến tự sự tuần tự. Các chiến lược này kết hợp cùng nhau nhằm đáp ứng các yêu cầu thể loại của video đánh giá ẩm thực, vốn cần cân bằng giữa kết nối cá nhân, độ tin cậy, sự rõ ràng mang tính giáo dục, tính dễ tiếp cận và mạch tự sự. Nghiên cứu đóng góp về mặt lý luận bằng cách chứng minh khả năng ứng dụng của phân tích đa phương thức dựa trên SFL đối với dữ liệu video, và về mặt thực tiễn bằng cách đưa ra các hiểu biết sâu sắc cho người sáng tạo nội dung về cách thức phối hợp đa phương thức có chủ đích giúp tăng cường sự tương tác của khán giả và sự trân trọng văn hóa.

Từ khóa: Phối hợp đa phương thức; ngôn từ; hình ảnh; cận ngôn từ; ẩm thực đường phố Việt Nam; việc thu hút người xem

1. Introduction

Vietnamese street food has achieved significant international recognition, evidenced by TasteAtlas listing 26 of Vietnam's street food dishes in the Top 100 Southeast Asian Street Foods (Tam Anh, 2025) and the cuisine's feature in the Netflix series Street Food: Asia (Nguyen Quy, 2019). Furthermore, popular dishes like Banh mi have been ranked among the top must-try street foods globally by major travel sites like Fodor's Travel (VnExpress, 2016). Understanding how this vibrant cuisine is represented in the international sphere is crucial, as it not only helps boost tourism but also actively shapes Vietnam's culinary identity among global audiences. This representation is increasingly mediated through video content, therefore making the medium itself a vital area for analysis.

Food is viewed not just as a necessity, but as a major source of entertainment and exploration. This rise in food content has driven a significant change in how consumers make purchasing and dining decisions as studies show that fifty-two percent of internet users actively seek information online when searching for recipes, restaurant reviews, and brand consumption experiences before making a purchase (Nielsen, 2014, cited in Sahelices-Pinto et al., 2017). Moreover, food reviews are among the most popular content categories online (Bi, 2018), with platforms like YouTube making vlogger recommendations and reviews easily accessible to a wide audience (Briliana et al., 2020).

While social media influencers and content creators can significantly boost Vietnam's tourism by swaying their followers' travel decisions, the medium they utilize - videos - possesses a multimodal nature that offers unique analytical advantages, particularly for studying nuanced cultural expressions like street food. With its inherent integration of many layers of meaning, video affords a unique lens for conducting a multimodal discourse analysis (MDA) of street food YouTube videos. MDA allows researchers to analyze various modes of communication (e.g., visual, verbal, paralinguistic, etc) simultaneously to understand how meaning is constructed and communicated. Researchers can determine which semiotic principles are common to all modes of meaning production and how these principles are implemented differently in each mode by examining meaning making across modes (Jewitt et al., 2025). Specifically, visual elements (e.g., video footage, images) can be examined to understand how they convey cultural values and sensory experiences of street food. Spoken language, on-screen text, and subtitles, on the other hand, can be investigated to help in understanding how language is used to describe food, narrate stories, and engage viewers. Other embodied modes such as the gestures, facial expressions, and body language of street food vendors or consumers can provide insights into cultural practices and social interactions surrounding street food. In terms of filmic modes, editing techniques, camera angles, and transitions can be analyzed to see how they contribute to the storytelling and overall engagement of the video. This includes looking at how close-ups of food, slow-motion shots and other cinematic techniques enhance the viewer's experience.

While previous research shows that MDA is a powerful tool for understanding how meaning is created in video, significant gaps remain. Existing studies on video reviews typically isolate aspects such as viewer engagement (Cunningham & Craig, 2017) or the interaction between linguistic features and visual cues at a certain static moment (Cenni & Vasquez, 2025) without systematically exploring how these elements work together to engage viewers into the video content.

With all the above-mentioned reasons, the study seeks answers to the research question "*How do verbal, visual and paralinguistic modes in a YouTube food review video orchestrate to engage viewers in appreciating Vietnamese street food?*" with the aim of uncovering the multimodal mechanisms through which viewers are persuaded to appreciate Vietnamese street food in video content.

2. Literature review

2.1 The relationship between interpersonal meaning and viewer engagement

There are several common approaches to studying videos from a MDA perspective, one of which is Systemic Functional Linguistics (SFL). While initially designed for the study of language by Halliday (1985), systemic functional theory is fundamentally a theory of meaning. As a result, its core principles extend beyond linguistics and can be applied to the analysis of other semiotic resources. Over time, SFL has been adapted and expanded to examine how both spoken and written language, along with non-linguistic resources such as images, gestures, spatial arrangements, 3D objects, sound, and music - contribute to meaning-making. SFL views language as a social semiotic system organized to perform three concurrent metafunctions: ideational (representing experiences), interpersonal (enacting social relations), and textual (organizing discourse).

Regarding the verbal mode, interpersonal meaning in SFL is primarily concerned with how language is used to establish, maintain and negotiate relationships between speakers and listeners (Matthiessen et al., 2010). It is about how people use language to take on roles, express their attitudes and values and negotiate social relations. Halliday describes interpersonal function as "language as action" (Halliday et al., 2004, p.29) to emphasize its role in the exchange of information and goods and services.

In terms of visual mode, interactive meaning in visual communication, which aligns with Halliday's interpersonal metafunction, refers to the relations established between the producers and the viewers of images (Kress & van Leeuwen, 2021). It concerns how images engage the viewer and position them in relation to the represented participants.

Concerning paralanguage, interpersonal meaning concerns how social relations are enacted and attitudes are expressed within discourse. Ngo et al. (2022) argue that paralanguage plays a critical role here by engaging with appraisal systems, particularly as its semovergent and sonovergent dimensions can resonate with interpersonal functions in language.

In the context of food review analysis, this interpersonal function becomes the central mechanism for establishing rapport and negotiating the acceptance of the subject matter. The host can actively use the resources of the interpersonal metafunction to perform the communicative action of taking on the role of the trusted expert and affective guide. Therefore, studying interpersonal meaning allows tracing the moments where the host strategically orchestrates persuasion across different modes to orient the viewer's attitude toward appreciating Vietnamese street food.

2.2 Realizations of viewer engagement through selected aspects of SFL-based models

Viewer engagement can be realized through selected aspects of SFL-based models, specifically through verbal, visual and paralinguistic resources that make meaning of the Interpersonal/Interactive metafunction.

Within the investigation of verbal resources, two key SFL-based frameworks can be utilized. The Mood system (Halliday et al., 2004) enables an analysis of the grammatical choices made by reviewers to enact their interactive roles and speech functions. By categorizing utterances as indicative (declarative statements that convey information or exclamative expressions of strong feeling), imperative (commands or suggestions) or interrogative (questions), the study can understand how reviewers construct their authority, invite participation, express enthusiasm or seek confirmation from the audience. Complementing this, Martin and White's Appraisal framework (2005) focuses on the Attitude system to unpack the evaluative language used. By identifying expressions of affect (emotional responses), judgement (evaluations of human behavior or character) and appreciation (evaluations of objects, processes, or phenomena), the investigation can reveal the ways reviewers express their subjective stance, build rapport with their audience and persuade them regarding the quality and experience of Vietnamese street food.

Within the visual resources, for the Interactive function, Kress & Leeuwen's (2021) equivalent term for interpersonal meaning, the investigation delves into Participants, Distance, Attitude, and Gaze. Analyzing the Participants helps to discern how the reviewer, the food and other individuals are visually presented in relation to the viewer, hence influencing perceived social connections. Distance is examined to understand the level of intimacy or objectivity established with the food and the audience, which impacts how personally involved the viewer feels. Attitude, which is conveyed through camera angle, reveals the reviewer's visual stance or evaluation of the food, ultimately contributing to the persuasive force of the review. Finally, Gaze is analyzed to understand how the reviewer visually engages or invites observation from the audience, directly shaping the interactive dynamics. By scrutinizing these interpersonal visual elements, the examination can uncover how the reviewer constructs rapport, expresses evaluation and encourages engagement with the audience regarding Vietnamese street food.

Within the domain of paralanguage resources, the investigation concentrates on Facial affect and Voice affect, aspects put forward by Ngo et al. (2022). Facial affect, encompassing expressions of joy, satisfaction, surprise, or distaste, directly communicates the reviewer's emotional response and evaluation of the food, which significantly influences the audience's perception and engagement. The areas of the face considered most significant are "eyebrows, eyes and mouth" (Ngo et al., 2022, p.120). Concurrently, Voice affect, including variations in tone, pitch, volume and speaking rate, conveys the reviewer's enthusiasm, sincerity or criticality to shape the interactive dynamic and persuasive power of the review. By analyzing these interpersonal paralanguage elements, the study can uncover how reviewers express their subjective experience, effectively connect with and persuade their viewers regarding the street food.

2.3 An adapted model for multimodal orchestration analysis

This study argues that viewer engagement is fundamentally driven by the host's strategic deployment of interpersonal meaning across multiple modes. Existing models (Barthes, 1977; Martinec & Salway, 2005; Royce, 1999; Unsworth, 2001) primarily deal with two-dimensional discourse, such as printed texts or static web pages where the main relationship is between text and a single fixed image. However, video content, like YouTube reviews, presents a multi-dimensional challenge. It involves the dynamic interplay of multiple modes simultaneously and sequentially. In a video, the visual content, verbal content and paralanguage are not static but flow continuously over time. This dynamic and time-based nature demands a more complex analytical approach that can account for the temporal and synergistic relationships between all these modes, beyond the traditional text-image paradigm. Accordingly, this section presents an adapted multimodal orchestration model specifically designed to map the relationships among these three modes at the moment of co-occurrence with a view to providing the specialized tool necessary for analyzing the host’s strategic engagement techniques.

***Verbal - Visual interaction**

Placed in Social Semiotics, van Leeuwen (2005) explored this interaction from the multimodal cohesion perspective. He examined how different semiotic resources, including language and images, integrate to form meaningful texts and communicative events. This approach bridges abstract possibilities for linking messages with Barthes’ (1977) original text-image relations to establish correspondences between their categories.

Table 1

Overview of visual–verbal interaction (van Leeuwen, 2005, p.230)

Image–Text Relations		
Elaboration	Specification	The image makes the text more specific (illustration)
		The text makes the image more specific (anchorage)
	Explanation	The text paraphrases the image (or vice versa)
Extension	Similarity	The content of the text is similar to that of the image
	Contrast	The content of the text contrasts with that of the image
	Complement	The content of the image adds further information to that of the text, and vice versa (“relay”)

To fully comprehend how narrative is constructed in videos, such as YouTube review videos about Vietnamese street food, it is crucial to analyze not only image-text relations but also information linking for both visual and verbal components. While van Leeuwen’s (2005) model is fundamental for understanding the immediate relationship between verbal commentary and visuals, it primarily describes what words and images do to each other in terms of meaning specification or complementarity at a given moment. This model, if solely used, is restricted to static texts, similar to previously mentioned frameworks.

Information linking delves into the cognitive connections between “bits” of information in order to reveal how they relate through categories like causal, temporal, or additive relationships, which are essential for creating meaningful sequences and narratives (ibid, 2005). In videos, especially those with a narrative flow like food reviews (e.g., showing preparation steps, tasting, and reactions), temporal linking is a dominant and fundamental aspect of storytelling. The sequence of "previous event," "next event," or "simultaneous event" links between shots, and between spoken words and corresponding visuals, enables the construction of a coherent story or procedural guide. Therefore, to fully grasp how a YouTube food review video constructs its narrative - from showcasing the ingredients and cooking process to detailing the tasting experience and offering a final recommendation - analysis of both the immediate image-text relations and the broader patterns of information linking is indispensable for understanding the content's coherence and progression. Information linking is realized through visual linking, including "Elaboration (Overview and Detail) and Extension (Temporal (Next event, Previous event and Simultaneous event), Spatial (Proximity and Co-presence) and Logical (Contrast and Similarity))" (van Leeuwen, 2005, p.229). It is further realized through verbal linking, namely "Elaboration (Explanation, Example, Specification, Summary and Correction) and Extension (Addition, Temporal, Spatial and Logical)" (van Leeuwen, 2005, p.225).

***Verbal - Paralanguage synergy**

According to Ngo et al. (2022), there are two primary ways paralanguage relates to spoken language: convergence or divergence. Paralanguage can semantically converge with verbal elements, meaning it aligns with the discourse semantics of spoken language. From an interpersonal perspective, it resonates with appraisal systems, expressing emotion through facial expressions, bodily stance, muscle tension and voice quality. Divergence, on the other hand, occurs when the verbal and paralinguistic modes provide contrasting meanings, which can be used strategically to achieve specific interactional effects.

***Paralanguage – Verbal - Visual orchestration**

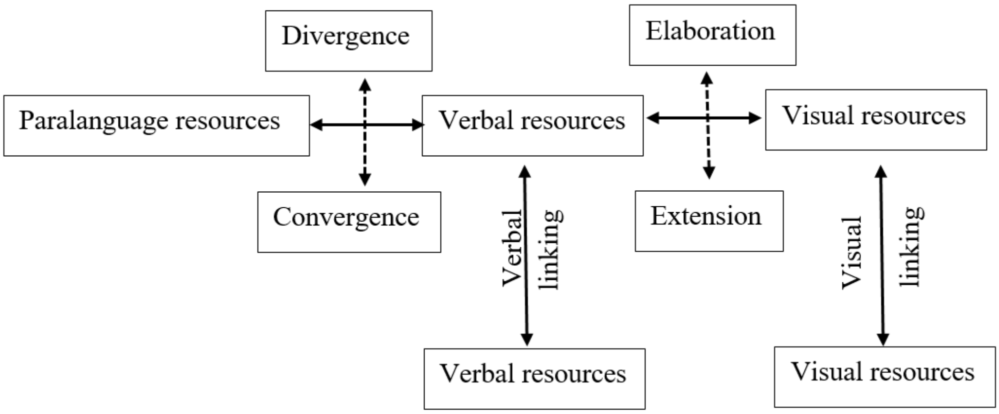
Based on previous arguments, the study proposes an adapted multimodal orchestration model to investigate the interaction among the three modes chosen for the exploration of engaging viewers to appreciate Vietnamese street food in review videos, based on the framework of Ngo et al. (2022) and van Leeuwen (2005).

Central to the model are the verbal resources which interact with both paralanguage and visual resources, realizing the "intersemiosis" relationship (O'Halloran, 2005, p.16) across three semiotic resources. Paralanguage can either diverge or converge with verbal elements whereas verbal and visual resources share a two-way collaboration namely elaboration and extension. In dynamic data like videos, verbal and visual resources also interact with themselves through a system of verbal and visual linking as meaning unfolds over time. Verbal linking and visual linking contribute to the "intrasemiosis" interaction (O'Halloran, 2005, p.16) – meaning within each semiotic resource. Therefore, meaning is achieved through both intersemiosis and intrasemiosis processes.

The adapted multimodal orchestration model of the study is visualized as in Figure 1:

Figure 1

The adapted multimodal orchestration model of the study



3. Methodology

The data of the study is the most viewed YouTube video exploring Vietnamese street food, which belongs to an English-speaking channel named "Best Ever Food Review Show". This channel distinguishes itself by pioneering a fresh and impactful approach to exploring global cuisines, which emphasizes the discovery of each country's distinctive foods through a dynamic presentation that fosters cultural empathy and understanding. Based in Vietnam, the host created the channel on September 23, 2010, and uploaded his first video in December 2015. As of June 2025, the channel boasts approximately 11.3 million subscribers. The video tells about a journey in which the hosts, who are identified as Sobe and Tien, discover local street food in Hanoi, including water bug extract, sandworm cakes, roasted quails with Vietnamese curry leaves, pig penal and others. The video, lasting 13.07 minutes, was uploaded in 2018, and as of July 2025, it records the view count of nearly 18 million times. There are 27 scenes in this video. The video can be viewed at https://www.youtube.com/watch?v=l5H6jl_g2XE&t=631s.

In this study, a scene is the most effective unit of analysis because it represents a complete piece of action in which there is no change in time or place (Iedema, 2001). Robert McKee, an Oscar-winning screenwriter, claims that "ideally, every scene is a story event" (1997, p. 56), elaborating on the complete meaning a scene can offer. Following that, Sánchez-Escalonilla (2014, p. 14, as cited in Figuero-Espadas, 2019) also concurs that the scene represents a "dramatic action unit" or an "event" where "something specific occurs". All of these features of the scene make it a crucial unit for MDA, as it is the smallest logical unit where all the modes - verbal, visual, and paralinguistic - are consistently collaborating to create a coherent segment of meaning. This makes a scene a more meaningful unit for uncovering the representations and identities being constructed than the more fragmented or larger-scale levels.

The data processing consists of the following steps.

Firstly, the scenes were transcribed and checked manually for accuracy. Multimodal transcription serves to reveal the co-deployment of semiotic resources and their dynamic unfolding over time (Baldry & Thibault, 2006). It involves making methodological decisions about which dimensions are key to the investigation as it can be a time-intensive and laborious process (Cenni & Vasquez, 2025). Hence, this study takes on Norris (2019)'s viewpoint that a selected number of key dimensions should be analyzed and discussed in-depth. The three selected modes are verbal, paralanguage and visual, as discussed previously. The multimodal transcription was stored in an Excel document for further analysis. Specifically, each mode was transcribed as follows:

- Verbal mode: Spoken dialogue and any voiceover were transcribed and manually checked for accuracy.

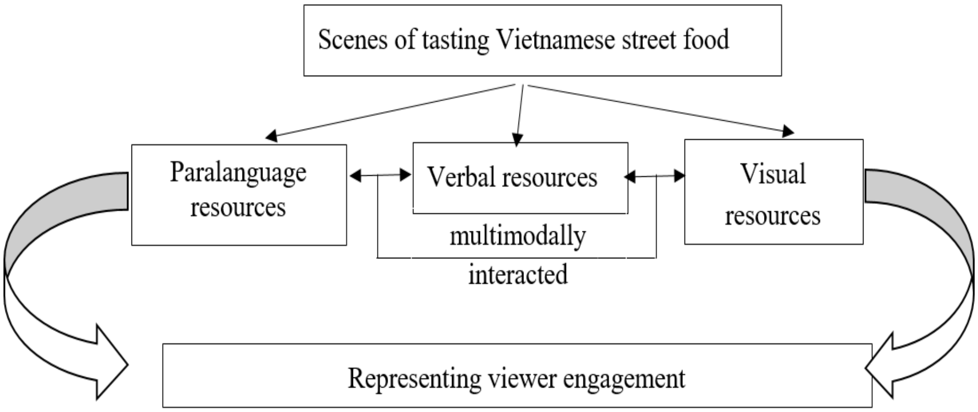
- Visual mode: Screenshots of the frames were taken, accompanied by a description of the actions and the setting of the images.

- Paralanguage mode: Hand/Arm movements, facial expressions, voice quality of the hosts/vendors were noted.

Secondly, the data were analyzed according to the proposed orchestration model as in Figure 1 and the analytical framework as illustrated in Figure 2.

Figure 2

The analytical framework of the study



In addition, this study acknowledges that the researcher's individual perspectives and experiences inherently influence the interpretation of the data. Therefore, several measures were implemented to ensure the credibility and transferability of the study, as these concepts are central to qualitative research (Paltridge & Phakiti, 2015). Firstly, prolonged engagement with the YouTube food review video genre was undertaken, involving extensive immersion in the content to develop a deep understanding of the online representations to avoid superficial observations (Creswell & Creswell, 2018). Secondly, triangulation was central, particularly through the use of multiple analytical perspectives (verbal, visual, and paralanguage analyses). As for transferability, which pertains to the applicability of the findings to other contexts, the study acknowledges that it does not aim for statistical generalization but rather a rich understanding of particular phenomena (Creswell & Creswell, 2018; Croker, 2009). Therefore, transferability was fostered by providing a rich and detailed description of the specific findings from the multimodal analysis.

4. Findings and discussion

4.1 Findings of the study

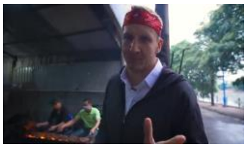
The analysis of the data revealed five multimodal strategies used to engage viewers in appreciating Vietnamese street food.

4.1.1 Creating interpersonal closeness through direct address

This subsection examines how reviewers foster a sense of interpersonal closeness by addressing viewers directly, using gaze, verbal address and affective tone to collapse the distance between speaker and audience. Through deliberate choices in mood, camera framing and paralinguistic cues, the hosts position viewers as co-participants in the tasting experience, thereby mirroring face-to-face interactions in addition to strengthening engagement and immediacy.

Table 2

Scene I.4th.2.(4) Sobe and Tien talk to the owner of the store

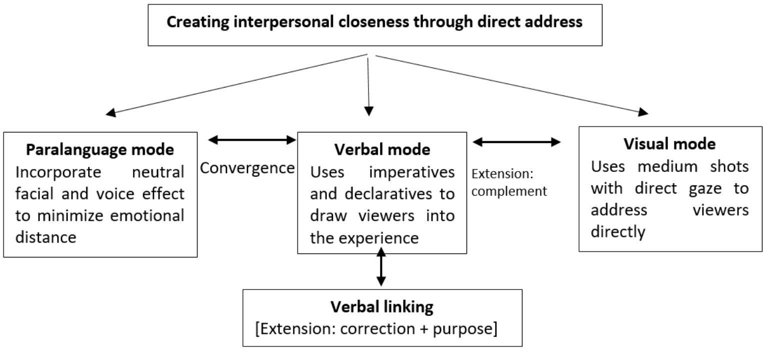
	Verbal mode	Paralanguage mode	Visual mode
Sobe	Hold up, guys, be mature, it's food, we can be adults about this. Okay, let's go learn about it.	- Face: blank expression - Voice: neutral tone	

In Scene I.4th.2.(4), when Sobe and Tien prepare to talk to the owner of the restaurant about the pig penal dish, Sobe engages viewers by directly addressing them as co-participants in the moment of encountering an unfamiliar dish, achieved through the orchestration of verbal, visual and paralinguistic resources. Verbally, the utterances consist of a sequence of imperatives and declaratives that create an [Extension: correction + purpose] intrasemiosis by moving from a call for composure to a forward-oriented invitation. The use of "guys" and "let's" constructs solidarity to draw the viewers into the experience as if sharing the challenge together. By using judgement-related terms (mature, adults), the imperative and declarative statements serve as a direct correction to the audience, through which potential negative feelings (such as disgust or immaturity) often associated with challenging dishes are managed. Visually, the medium shot of Sobe speaking directly toward the camera, with the vendor grilling food in the background, establishes a balance between intimacy and context. The frontal angle and eye-level shot position him and the audience as equals, which encourages involvement through a conversational stance. This visual composition offers an [Extension: complement] intersemiotic alignment with the verbal invitation to "learn about it," as the viewer simultaneously sees the food being prepared, thus reinforcing the sense of joining Sobe on-site. Paralinguistically, Sobe's neutral tone and blank facial expression minimize emotional distance. This facial and voice effect allows the verbal mode to take prominence in guiding viewer alignment rather than dramatizing fear or excitement. The unbiased paralanguage ensures that the viewer's focus remains on the inclusive appeal of the speech and the inviting visual frame. Together, these modes work intersemiotically to frame a direct encounter with viewers.

Accordingly, the mechanism of creating interpersonal closeness through direct address is contextualized with variable realizations. The mechanism can be visualized as in Figure 3.

Figure 3

The mechanism of creating interpersonal closeness through direct address


Creating interpersonal closeness through direct address
Paralanguage mode
Incorporate neutral facial and voice effect to minimize emotional distance
Convergence
Verbal mode
Uses imperatives and declaratives to draw viewers into the experience
Extension: complement
Visual mode
Uses medium shots with direct gaze to address viewers directly
Verbal linking
[Extension: correction + purpose]

4.1.2 Claiming expert authority through objective attitude

This subsection focuses on how reviewers enhance trustworthiness by adopting an objective, matter-of-fact evaluative stance, using descriptive verbal appraisal, steady vocal delivery and visually neutral framing to project composure. Through this multimodal performance of expertise, the reviewers position themselves not merely as enthusiastic consumers but as knowledgeable evaluators whose judgements viewers can rely on.

Table 3

Scene I.1st.9.(10) Sobe and Tien taste the waterbug

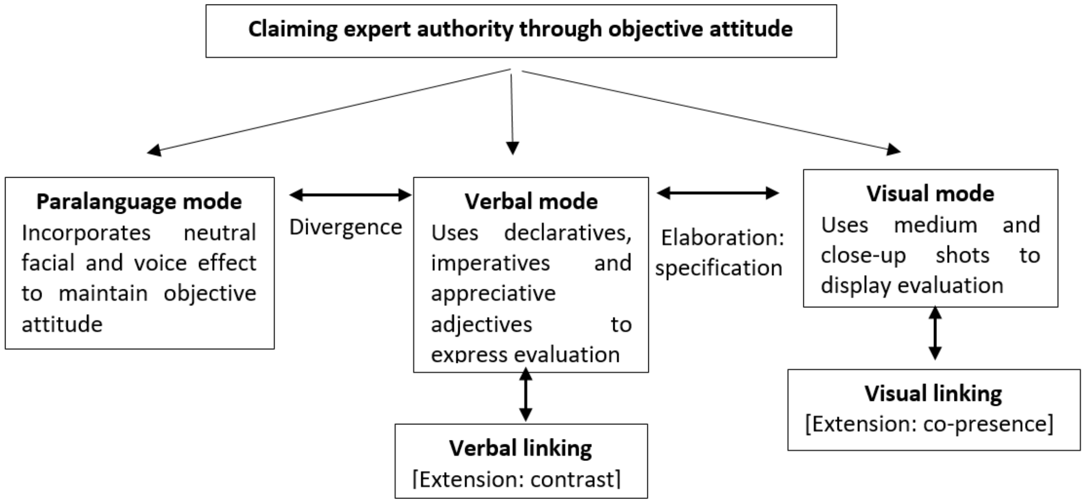
	Verbal mode	Paralanguage mode	Visual mode
Sobe	Yo,		
Sobe	the body feels so disgusting,		
Sobe	but the flavor permeating out, it's amazing.	- Face: Sobe (neutral); Tien (smiling); Sobe looking at Tien - Voice: emphasis on "amazing"	
Tien	Very special.	- Face: looking at the camera and smiling - Voice: spirited tone	

In Scene I.1st.9.(10) when Sobe and Tien taste the waterbug, expert authority is realized through a complex interplay of verbal, visual and paralinguistic resources, where honesty emerges not from uniform positivity but from a multi-layered evaluative stance. Verbally, Sobe's declarative "Yo," "the body feels so disgusting," followed by "but the flavor permeating out, it's amazing", reveal an [Extension: contrast] intrasemiosis within the verbal mode. This willingness to articulate the negative aspect ("disgusting") first, even when the overall outcome is positive, signals that the host is a meticulous evaluator who refuses to ignore a flaw. Tien's brief appreciative comment "Very special" further stabilizes the judgement by providing a culturally respectful framing. Paralinguistically, however, Sobe's neutral facial expression when saying "it's amazing" creates a divergence between verbal enthusiasm and emotional display, which paradoxically enhances trust. The lack of exaggerated facial affect signals that the appraisal is reflective rather than performative. The upward vocal emphasis on "amazing," and Tien's gentle smile and spirit-up tone when saying "Very special," work [divergently] with the verbal mode, giving evidence for their mixed reactions, from hesitation and surprise to eventual appreciation. Visually, the medium shot of both hosts slowly placing the waterbug into their mouths situates their evaluation within an authentic moment of sensory negotiation. The cut to a close-up of the waterbug being sliced during the utterance "the body feels so disgusting" performs an [Elaboration: specification] verbal–visual intersemiosis to visually substantiate the discomfort described verbally. When the shot returns to the two hosts, Tien looks directly at the camera while saying "very special". This creates a moment of [Extension: co-presence] visual intrasemiosis as his gaze shifts meaning from peer-to-peer reaction to a viewer-addressing endorsement. Collectively, these multimodal choices construct trustworthiness by showing that the hosts' responses are layered, candid and grounded in actual sensory experience.

Accordingly, the mechanism of claiming expert authority through objective attitude can be visualized as in Figure 4.

Figure 4

The mechanism of claiming expert authority through objective attitude


Claiming expert authority through objective attitude
Paralanguage mode
Incorporates neutral facial and voice effect to maintain objective attitude
Divergence
Verbal mode
Uses declaratives, imperatives and appreciative adjectives to express evaluation
Elaboration: specification
Visual mode
Uses medium and close-up shots to display evaluation
Verbal linking
[Extension: contrast]
Visual linking
[Extension: co-presence]

4.1.3 Explaining food values through logical reasoning

This subsection examines how reviewers engage viewers by offering clear and rational explanations for why a dish is worth appreciating, using factual descriptions, nutritional reasoning, and process-based explanations to legitimise their positive evaluations. By grounding appreciation in logical verbal argumentation and visually supported evidence, reviewers construct a persuasive rationale that moves beyond personal taste.

Table 4

Scene I.2nd.3.(7) Sobe and Tien talk about the sandworms and even touch the living sandworms

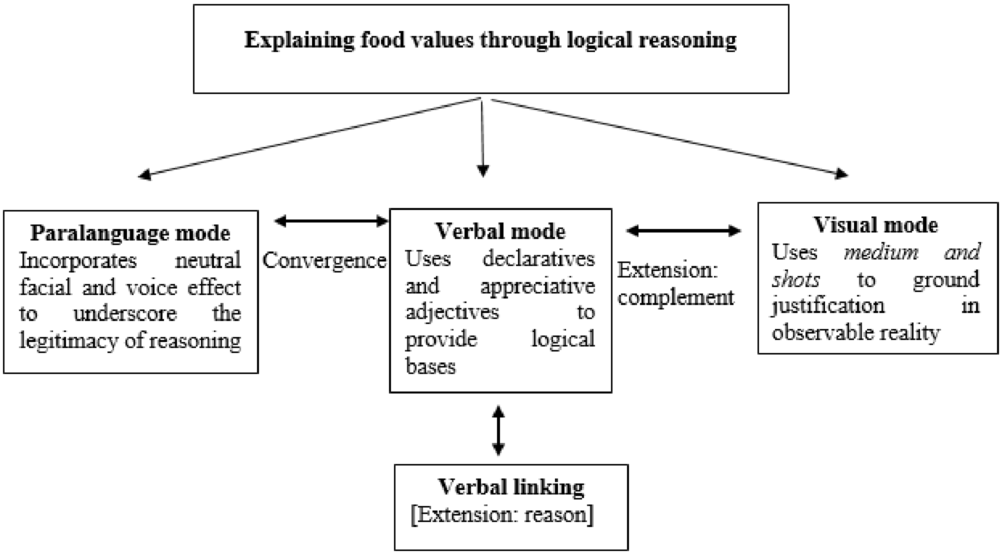
	Verbal mode	Paralanguage mode	Visual mode
Sobe	Um, we're making pancakes with worms. It is the best way to get some protein in your diet and also still be able to eat delicious pancakes.	- Face: neutral - Voice: emphasis on "protein"	

In Scene I.2nd.3.(7) when Sobe and Tien talk about the sandworms and even touch the living sandworms, the host engages viewers by shifting from affective reactions to a more reasoned framing of the dish, thereby strengthening the credibility of their appreciation. Verbally, Sobe's declaratives illustrate intrasemiosis within the verbal mode as he moves from a factual description ("making pancakes with worms") to a rational justification grounded in nutritional value ("protein") and pleasure value ("delicious pancakes") with an appreciative adjective. This [Extension: reason] progression from observation to explanation provides a logical basis for appreciating the dish. Paralinguistically, his neutral facial expression and vocal emphasis on "protein" underscore the legitimacy of his reasoning, working [convergently] with the verbal mode to highlight the nutritional argument as the key point viewers should attend to. Visually, the medium shot of Sobe holding a live sandworm in his palm, with the vendor preparing the dish in the background, creates an [Extension: complement] verbal–visual intersemiosis. Particularly, the viewer sees the ingredient and its preparation context at the same moment Sobe explains its dietary value, which helps to ground his justification in observable reality. Together, these multimodal resources construct a rational basis for appreciation by aligning explanation, evidence and embodied demonstration in order for viewers to understand *why* the dish merits respect beyond its initial shock value.

Accordingly, the mechanism of explaining food values through logical reasoning can be visualized in Figure 5.

Figure 5

The mechanism of explaining food values through logical reasoning


Explaining food values through logical reasoning
Paralanguage mode
Incorporates neutral facial and voice effect to underscore the legitimacy of reasoning
Convergence
Verbal mode
Uses declaratives and appreciative adjectives to provide logical bases
Extension: complement
Visual mode
Uses *medium and shots* to ground justification in observable reality
Verbal linking
[Extension: reason]

4.1.4 Using familiar comparisons to reframe the unfamiliar

This subsection examines how reviewers further facilitate viewer acceptance by drawing analogies between unfamiliar Vietnamese street foods and more easily relatable food items. By verbally mapping new sensory experiences onto established culinary categories and aligning these comparisons with visual and paralinguistic evidence, reviewers reframe the unfamiliar as recognizably familiar, which effectively lowers perceived barriers of strangeness.

Table 5

Scene I.4th.4.(4) Sobe and Tien taste the grilled urethra and give comments

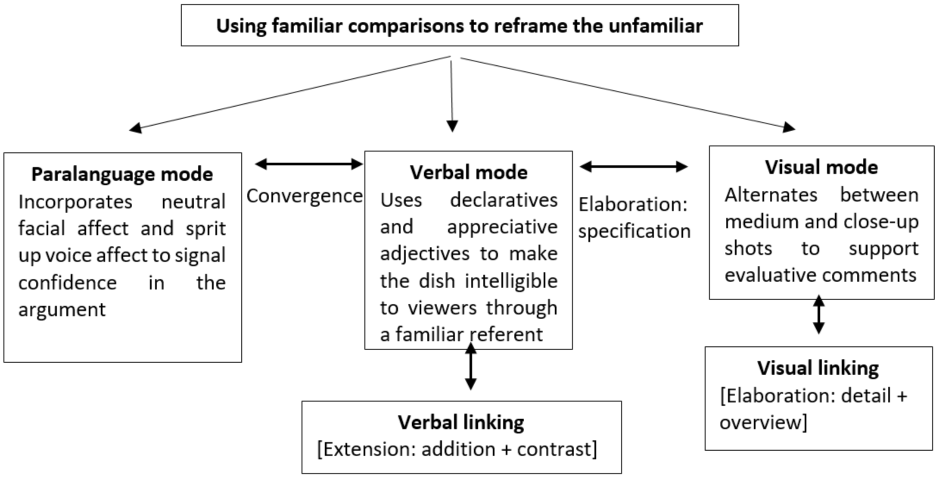
	Verbal mode	Paralanguage mode	Visual mode
Sobe	It's like a massive piece of cartilage,	- Face: neutral - Voice: speaking quickly (26 syllables in 3 seconds, approximately 9 syllables per second)	
Sobe	really crunchy. It reminds me a little bit of chicken feet.		
Tien	Yep.		
Sobe	But it's easier to get through.		

Scene I.4th.4.(4) shows Sobe and Tien tasting the grilled urethra and giving comments. In this scene, the reviewers help convert potential viewer hesitation into openness by comparing the grilled urethra to familiar foods with a view to making the dish cognitively accessible. Verbally, Sobe's series of declaratives illustrates an [Extension: addition + contrast] intrasemiosis within the verbal mode, moving from physical description ("massive piece of cartilage," "crunchy") to analogy ("chicken feet") and finally to evaluation ("easier to get through"). This internal shift from description to comparison then to judgement provides viewers with a rational bridge between the unfamiliar and the familiar. His comparison to chicken feet, a more common texture for many global audiences, uses an appreciative adjective ("easier") within the Appraisal framework to normalize the dish and ease viewer apprehension. Paralinguistically, Sobe's neutral facial expression and fast speech tempo (approximately 9 syllables per second) signal confidence and lack of hesitation, reinforcing the credibility of his comparison. This paralinguistic steadiness works [convergently] with the verbal mode. In other words, the quick and fluid delivery supports the sense that the analogy is natural and trustworthy rather than forced. Visually, the medium shot of the two hosts eating, followed by a close-up of the urethra on the grill and then a return to the eating scene, creates an [Elaboration: specification] verbal–visual intersemiosis, allowing viewers to verify the textual qualities ("cartilage," "crunchy") against the actual appearance of the food. Within the visual mode, an [Elaboration: detail + overview] intrasemiosis unfolds as the camera alternates between hosts and food, linking their evaluative comments directly to visual evidence and reinforcing the relational closeness between taster and dish. Collectively, these multimodal resources work to engineer conversion: verbal analogies make the dish relatable, paralinguistic cues convey confidence, and visual confirmation grounds the comparison in observable reality. This orchestration ultimately persuades viewers to reconsider an initially challenging food.

Accordingly, the mechanism of using familiar comparisons to reframe the unfamiliar can be visualized in Figure 6.

Figure 6

The mechanism of using familiar comparisons to reframe the unfamiliar



Using familiar comparisons to reframe the unfamiliar
Paralanguage mode
Incorporates neutral facial affect and sprit up voice affect to signal confidence in the argument
Convergence
Verbal mode
Uses declaratives and appreciative adjectives to make the dish intelligible to viewers through a familiar referent
Elaboration: specification
Visual mode
Alternates between medium and close-up shots to support evaluative comments
Verbal linking
[Extension: addition + contrast]
Visual linking
[Elaboration: detail + overview]

4.1.5 Pacing the narrative through sequential progression

This subsection examines how reviewers engage viewers by guiding them through a step-by-step unfolding of events, using verbal commentary, paralinguistic cues and temporally ordered visual editing to construct a coherent narrative flow. Through this sequential progression, viewers experience the discovery of Vietnamese street food as a developing storyline, thus enhancing anticipation and sustaining attention.

Table 6

Scene I.1st.8.(10) Sobe and Tien watch the waitress cut the waterbug

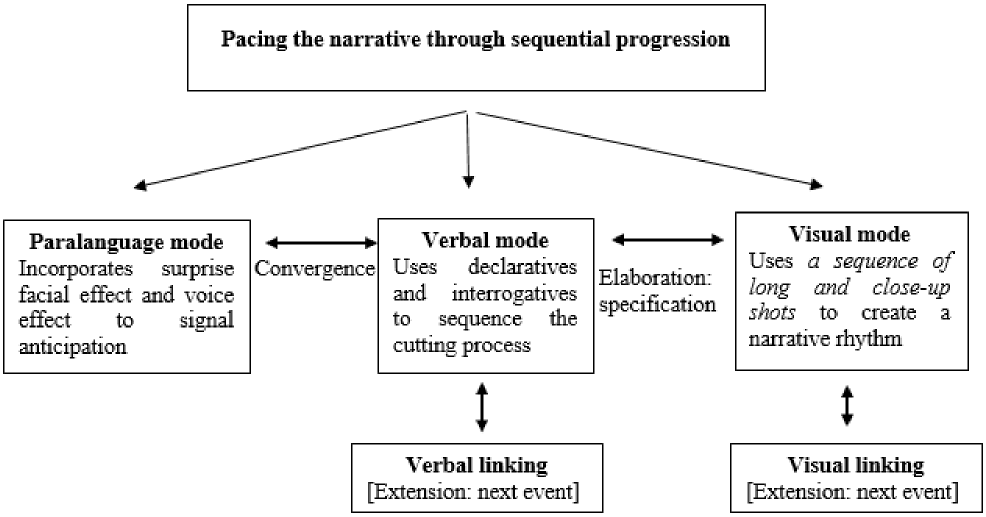
Verbal mode		Paralanguage mode	Visual mode
Sobe	Oh she's gonna do it here?	- Voice: raised tone on "here"	
Tien	Uh, gonna do, cut, and that's the head.		
Sobe	Huh … She puts, she cuts the head off right into the sauce,	Voice: prolonged and lowered "huh"	
Sobe	and then the body, and then ..hum...that's called a thorax.	- Voice: raised tone on "body"; lowered tone on "hum"	
Tien	Right.		
Sobe	I know that, for sure, and then she walked away. What just happened?	- Face: surprised expression - Voice: raised tone on "happened"	

Scene I.1st.8.(10) shows Sobe and Tien watching the waitress cut the waterbug. In this scene, temporal extension structures a dynamic unfolding of action that draws viewers into the moment-by-moment sequence of the waterbug preparation. Verbally, Sobe narrates each stage of the cutting process through a series of declarative clauses ("She's gonna do it here?", "that's the head", "she cuts the head off… and then the body"), creating a running commentary that mirrors the chronological steps of the action. His use of the interrogative ("What just happened?") further heightens suspense and signals a shared discovery with viewers. Paralanguage reinforces this sense of unfolding anticipation. That is, the raised pitch on "here" conveys surprise at the immediacy of the cutting while the prolonged and lowered "huh" marks hesitation and uncertainty. Moreover, his facial expression of surprise aligns [convergently] with these vocal cues, giving viewers embodied access to his reaction. Visually, the alternating medium shots of the hosts watching and close-up shots of the waterbug being picked up and segmented construct a clear temporal chain. This visual sequencing creates a narrative rhythm, from observation to action and reaction. An [Elaboration: specification] intersemiosis emerges as verbal narration synchronizes with specific visual stages of the cutting process, which results in a tightly integrated multimodal storyline where speech explains what the viewer sees and visuals confirm what is being told. Simultaneously, paralanguage–verbal intersemiosis amplifies affective shifts: raised intonation marks moments of unexpectedness which is visible in the close-ups while lowered tones coincide with visually unsettling actions. An [Extension: next event] intrasemiosis within the verbal mode appears through cohesive lexical repetition ("and then… and then… and then"), which mirrors the sequential cutting seen visually. Meanwhile, an [Extension: next event] visual intrasemiosis arises from repeated close-ups that foreground the same bug and bowl to maintain narrative continuity. The scene reaches a climax when Sobe turns to the camera and asks, "What just happened?", a direct gaze that transitions viewers from observers to participants in the unfolding narrative. Through this multimodal orchestration, the scene constructs narration not merely by telling a story but by staging it temporally and visually for the viewer to experience in real time.

Accordingly, the mechanism of pacing the narrative through sequential progression can be visualized in Figure 7.

Figure 7

The mechanism of pacing the narrative through sequential progression



Pacing the narrative through sequential progression

Paralanguage mode
Incorporates surprise facial effect and voice effect to signal anticipation

Convergence

Verbal mode
Uses declaratives and interrogatives to sequence the cutting process

Elaboration: specification

Visual mode
Uses *a sequence of long and close-up shots* to create a narrative rhythm

Verbal linking
[Extension: next event]

Visual linking
[Extension: next event]

4.2 Discussion of the findings

Upon analyzing how the three modes interact to construe meanings, it becomes evident that these modes do not operate in isolation but work together to create a compelling and immersive experience for the audience. The systematic application of the interpersonal metafunction uncovered five core engagement strategies embedded in the multimodal text: the creation of interpersonal closeness (direct address), the establishment of expert authority (objective attitude), the justification of food values (logical reasoning), the cognitive shift through reframing the unfamiliar (familiar comparisons) and the structural control of the narrative progression (sequential pacing). The results of this analysis illuminate how various multimodal resources complement or counteract to realize meanings in dynamic data like videos, which aligns with prior research (Lu, 2024; Vo, 2025; Zhang, 2022).

Specifically, the findings confirm the principle of "intersemiosis" (O'Halloran, 2005, p.16) - the active interplay among the three modes in the enactment of interpersonal meaning. In the analyzed scenes, the hosts' relationship with their audiences is achieved through specific forms of orchestration. Via *paralanguage-verbal convergence*, the credibility is gained when interpersonal meaning is most powerfully constructed through the synchronized use of the paralanguage and verbal modes. When a host's paralanguage (a fast-paced, high-pitched voice or an eager facial expression) aligns precisely with their verbal declarative (a positive appraisal) the message achieves congruence, which signals to the viewer that the host's excitement is genuine. This is critical for building the trust necessary for the viewer to accept the representation of the food. The trust is also achieved when a detailed verbal description of the food's qualities is immediately elaborated by a visual close-up of the specific ingredients or textures. The visual mode validates the verbal claim to ensure that the hosts' interpersonal appeal is grounded in demonstrable evidence. This leads to the overall representation of the street food as persuasive. Furthermore, *the intersemiosis between the verbal and visual modes* creates a sense of shared experience. For instance, a verbal declarative is complemented by a visual of the host looking into the camera (a "demand" gaze). This orchestration draws the viewer into the scene, turning a passive observation into a virtual interaction. This technique effectively closes the social distance between the reviewer and the viewer, which makes the appreciation of the street food a shared social act.

Unlike previous studies which mostly focus on the interaction of modes in their static form, the application of the proposed multimodal orchestration model reveals that interpersonal meaning of the representation is sustained and made persuasive by how the modes connect information over the scene's duration. The principle of "intrasemiosis" (O'Halloran, 2005, p.16) – how each resource interacts with itself over time - is crucial for achieving sustained viewer engagement. Within the *verbal mode*, reviewers frequently construct internally coherent evaluative chains through declarative and imperative structures, or positive/negative appreciative adjectives. The [Elaboration] and [Extension] relations help to realize the sequencing of appraisal (moving from surprise to positive evaluation and intensified appreciation), which guides viewers' expectations. Similarly, the *visual mode* exhibits its own internal coherence through its [Elaboration] and [Extension] links, in particular recurring shot types such as medium shots of the hosts eating, close-ups of the dish or consistent use of eye-level shots create a unified visual framing that maintains relational intimacy and continuity across turns. In this way, intrasemiosis provides the structural backbone upon which intersemiosis can operate, thus enabling verbal, visual and paralinguistic resources to align more effectively in generating viewer involvement, trust and appreciation. Through the combined action of these intrasemiotic mechanisms, the hosts ensure that the engagement strategy is not a static claim, but a dynamically unfolding and sustained persuasive journey.

Moreover, the multimodal strategies used to engage viewers are deeply shaped by the inherent affordances and expectations of the food review video format. Reviews, in general, and YouTube food reviews, in particular, function as hybrid genres that blend entertainment, information and personal storytelling (Cheng, 2023; Hyland & Diani, 2009; Kathpalia, 2021; Truong et al., 2025). Therefore, this genre requires presenters to maintain an ongoing interpersonal connection with a dispersed audience who cannot respond in real time.

The *first* mechanism shows that the host actively counteracts the viewer's physical absence through a core multimodal strategy: simulating conversational immediacy, which is also documented in Cenni and Vasquez's (2025) study. This involvement is achieved by orchestrating the verbal imperative mood (direct address) and the neutral voice/facial effect with the visual mode's consistent use of direct gaze and eye-level camera positioning. This strategic convergence is central to establishing Parasocial Interaction (Horton & Wohl, 1956), which approximates reality and creates the "illusion of intimacy at a distance" (Martin & Ballentine, 2005, p.198). Through these simulated personal encounters, viewers develop a sense of relational closeness with the host, gradually viewing the persona as a trusted peer. *Secondly*, the display of expert authority through objective tone reinforces the reviewer's position as a credible evaluator, a practice consistent with Pfeuffer and Phua's (2021) study, in which their findings indicate that reviewers should present themselves as knowledgeable to sustain viewer trust in review genres. The reviewers in this study echo this pattern by employing neutral facial/voice affect, appreciative language and eye-level shots, thereby meeting genre expectations for reliability. Importantly, this objective stance does not signal an absence of evaluation but reflects a controlled mode of presenting evaluation, where appraisal is framed as descriptive and object-focused rather than subjectively driven. Through the divergence between verbal appraisal and restrained paralanguage, reviewers are able to maintain credibility while still guiding viewers' assessments of Vietnamese street food. *Thirdly*, the use of logical reasoning about ingredients, cooking processes, or nutritional value reflects YouTube's educational function, which is a prominent affordance that YouTube can offer to users (Calude, 2023; Mostafa et al., 2023). By embedding brief explanations within the introduction sequence, reviewers provide viewers with accessible knowledge, allowing appreciation to emerge from understanding rather than passive consumption. *Fourthly*, the deployment of familiar comparisons to reframe the unfamiliar is especially important in the context of Vietnamese street food, where many global viewers may lack prior cultural or culinary familiarity. As noted by Avieli (2011), certain Vietnamese food items, particularly insects, are identified as exotic and even repulsive to non-Vietnamese audiences, since they tap into Western taboos against insect-eating. Therefore, a global audience can benefit from analogies that anchor new sensory experiences in recognizable categories. *Finally*, sequential narrative pacing mirrors the conventional structure of food review videos, which Briliana et al. (2020) describe as progressing through predictable stages, from introduction to preparation to tasting to evaluation, to facilitate cognitive coherence and viewing satisfaction. The sequencing choices in this dataset adhere to this template, enabling viewers to follow an interpretive journey that moves steadily from curiosity toward appreciation. All in all, the multimodal strategies identified in this study resonate strongly with patterns established in prior research, thus demonstrating how the food review video format requires a balanced orchestration of personal connection, credibility, accessibility and narrative flow.

5. Conclusion

This study sets out to determine how verbal, visual, and paralanguage modes orchestrate to engage viewers in appreciating Vietnamese street food. Through a multimodal discourse analysis of the most viewed food review video, the findings show that meaning is not produced by any single mode but through flexible orchestration involving both intersemiosis - the integration of meaning across modes - and intrasemiosis - the elaboration of meaning within each mode. To foster viewer engagement, reviewers employ multimodal mechanisms such as creating interpersonal closeness, claiming expert authority, explaining food values, using familiar comparisons and constructing sequential progression. These five strategies collectively fulfill the communicative demands of the food review video genre, which necessitates the synthesis of personal connection, critical credibility, educational understanding, cognitive accessibility and cohesive narrative flow. Overall, the study demonstrates that multimodal orchestration is both purposeful and adaptive in shaping how Vietnamese street food becomes narratively compelling and culturally appreciable to global audiences.

This study offers significant theoretical and practical contributions. Theoretically, the research validates the application of a systematic SFL-derived multimodal orchestration framework for the analysis of video data to prove its efficacy in revealing complex interpersonal meanings. Practically, the findings provide actionable insights for content creators, suggesting that effective viewer engagement and cultural appreciation rely on the deliberate synchronization of all semiotic resources rather than on any single mode.

Despite the analytical depth achieved, a primary limitation stems from the small sample size chosen for intensive qualitative and demonstrable multimodal analysis. While this approach guarantees rigorous detail, the findings related to the specific manifestations of intersemiosis cannot be immediately generalized across the entire corpus of YouTube food reviews. Future research should apply this orchestration framework to a larger and more diverse dataset to confirm the consistency of these persuasive strategies and enhance the external validity of the model.

6. Acknowledgment of AI Assistance

In the preparation of this manuscript, an artificial intelligence tool was employed exclusively for proofreading purposes, including grammar checking and language refinement. All research ideas, analyses and interpretations are entirely conducted by the authors.

Acknowledgement: *This study is funded by Hanoi National University of Education under the university-level scientific research project, code SPHN25-27.*

References

Avieli, N. (2011). Making sense of the Vietnamese cuisine. *Education about Asia*, *16*(3), 42-45. https://www.asianstudies.org/wp-content/uploads/making-sense-of-vietnamese-cuisine.pdf?utm

Baldry, A., & Thibault, P. (2006). *Multimodal transcription and text analysis*. Equinox.

Barthes, R. (1977). *Image, music, text*. Fontana.

Bi, N. C. (2018). Product review videos on YouTube as eWOM. In L. Ha (Ed.), *The audience and business of YouTube and online videos* (pp. 59–72). Rowman & Littlefield. https://www.bloomsbury.com/us/audience-and-business-of-youtube-and-online-videos-9781978755772/?

Briliana, V., Ruswidiono, W., & Deitiana, T. (2020). Do millennials believe in food vlogger reviews? A study of food vlogs as a source of information. *Journal of Management and Marketing Review*, *5*(3) 170-178. https://www.researchgate.net/publication/347381879_Do_Millennials_believe_in_food_vlogger_reviews_A_study_of_food_vlogs_as_a_source_of_information

Calude, A. S. (2023). *The linguistics of social media: An introduction* (1st ed.). Routledge. https://www.routledge.com/The-Linguistics-of-Social-Media-An-Introduction/Calude/p/book/9781032330945

Cenni, I., & Vásquez, C. (2025). Italian food experiences on Airbnb: A multimodal analysis of hosts' promotional videos. *Ibérica,* (49), 45-76. https://doi.org/10.17398/2340-2784.49.45

Cheng, S. (2023). A review of interpersonal metafunction studies in systemic functional linguistics (2012–2022). *Journal of World Languages*, *10*(3), 623-667. https://doi.org/10.1515/jwl-2023-0026

Creswell, J.W. & Creswell, J.D. (2018*). Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.) Sage.

Croker, R. A. (2009). An introduction to qualitative research. In J. Heigham & R. A. Crocker (Eds.). *Qualitative research in applied linguistics: A practical introduction*. (pp.3-24). Palgrave Macmillan. https://doi.org/10.1057/9780230239517_1

Cunningham, S., & Craig, D. (2017). Being 'really real' on YouTube: Authenticity, community and brand culture in social media entertainment. *Media International Australia, 164*(1), 71–81. https://doi.org/10.1177/1329878X17709098

Figuero-Espadas, J. (2019). A review of scene and sequence concepts. *Communication & Society*, *32*(1), 267-277. https://doi.org/10.15581/003.32.37829

Halliday, M. A. K. (1985). *An introduction to functional grammar*. Edward Arnold. https://www.cambridge.org/core/journals/studies-in-second-language-acquisition/article/abs/an-introduction-to-functional-grammar-michael-a-k-halliday-london-edward-arnold-1985-pp-384/0B10E76178E3B17CC5418DE0E8117C32

Halliday, M.A.K., Matthiessen, C.M.I.M., Halliday, M., & Matthiessen, C. (2004). *An introduction to functional grammar* (3rd ed.). Routledge. https://doi.org/10.4324/9780203783771

Horton, D., & Wohl, R.R. (1956). Mass communication and para-social interaction: Observations on intimacy at a distance, *Psychiatry*, *19*(3), 215-229. https://doi.org/10.1080/00332747.1956.11023049

Hyland, K., & Diani, G. (2009). Introduction: Academic evaluation and review genres. In K. Hyland & G. Diani (Eds.) *Academic evaluation* (pp. 1-14). Palgrave Macmillan. https://doi.org/10.1057/9780230244290_1

Iedema, R. (2001). Analyzing film and television: A social semiotic account of Hospital – an unhealthy business. In van Leeuwen, T. & Jewitt. C. (Eds.), *Handbook of visual analysis* (pp. 183–206). SAGE. https://doi.org/10.4135/9780857020062.n9

Jewitt, C., Bezemer, J., & O'Halloran, K. (2025). *Introducing multimodality* (2nd ed.). Routledge. https://www.routledge.com/Introducing-Multimodality/Jewitt-Bezemer-OHalloran/p/book/9781032845388

Kathpalia, S. S. (2021). *Persuasive genres: Old and new media* (1st ed.). Routledge. https://doi.org/10.4324/9780429243721

Kress, G., & van Leeuwen, T. (2021). *Reading images: The grammar of visual design* (3rd edition). Routledge. https://www.routledge.com/Reading-Images-The-Grammar-of-Visual-Design/Kress-vanLeeuwen/p/book/9780415672573

Lu, Y. (2024). Multimodal discourse analysis of the promotional film countdown: Beginning of spring for the opening ceremony of Beijing Winter Olympics. *International Journal of Language, Literature and Culture, 4*(4), 17–29. https://doi.org/10.22161/ijllc.4.4.3

Martin, B., & Ballantine, P. W. (2005). Forming parasocial relationships in online communities. *Advances in Consumer Research*, *13*(2), 197-202. https://researchportal.bath.ac.uk/en/publications/forming-parasocial-relationships-in-online-communities/

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. Palgrave Macmillan.

Martinec, R., & Salway, A. (2005). A system for image–text relations in new (and old) media. *Visual Communication*, *4*(3), 337-371. https://doi.org/10.1177/1470357205055

Matthiessen, C., Lam, M. & Teruya, K. (2010). *Key terms in systemic functional linguistics* (1st ed.). Bloomsbury. https://www.bloomsbury.com/us/key-terms-in-systemic-functional-linguistics-9781847064400/

McKee, R. (1997). *Story: Substance, structure, style, and the principles of screenwriting*. Harper-Collins Publishers.

Mostafa, M. M., Feizollah, A., & Anuar, N. B. (2023). Fifteen years of YouTube scholarly research: Knowledge structure, collaborative networks, and trending topics. *Multimedia Tools Application, 82*, 12423–12443. https://doi.org/10.1007/s11042-022-13908-7

Ngo, T., Hood, S., Martin, J. R., Painter, C., Smith, B. A., & Zappavigna, M. (2022). *Modelling paralanguage using systemic functional semiotics: Theory and application*. Bloomsbury. https://www.bloomsbury.com/au/modelling-paralanguage-using-systemic-functional-semiotics-9781350074910/

Norris, S. (2019). *Systematically working with multimodal data: Research methods in multimodal discourse analysis.* Wiley-Blackwell. DOI:10.1002/9781119168355

O'Halloran, K. L. (2005). *Mathematical discourse: Language, symbolism and visual images.* London: Continuum.

Paltridge, B., & Phakiti, A. (2015). *Research methods in applied linguistics: A practical resource* (2nd ed.), Bloomsbury. https://www.bloomsbury.com/uk/research-methods-in-applied-linguistics-9781472 524560/

Pfeuffer, A., & Phua, J. (2021). Stranger danger? Cue-based trust in online consumer product review videos. *International Journal of Consumer Studies, 46*(3), 964-983. https://doi.org/10.1111 /ijcs.12740

Royce, T. (1999). *Visual-verbal intersemiotic complementarity in the Economist magazine.* [Ph.D. Dissertation]. The University of Reading. http://www.isfla.org/Systemics/Print/Theses/Royce Thesis/

Sahelices-Pinto, C., Lanero-Carrizo, A., Vázquez-Burguete, J. L., & Gutierrez-Rodriguez, P. (2018). Ewom and 2.0 opinion leaders in the food context: A study with a sample of Spanish food-related weblogs. *Journal of Food Products Marketing*, *24*(3), 328–347. doi:10.1080/10454446 .2017.1266561

Truong, T. A., Piscarac, D., Kang, S. M., & Yoo, S. C. (2025). Virtual culinary Influence: Investigating the impact of food vlogs on viewer attitudes and restaurant visit intentions. *Information*, *16*(1), 44. https://www.mdpi.com/2078-2489/16/1/44

Unsworth, L. (2001). T*eaching multiliteracies across the curriculum: Changing contexts of text and image in classroom practice.* Open University Press.

van Leeuwen, T. (2005). *Introducing social semiotics*: *An introductory textbook*. Routledge. https://www.routledge.com/Introducing-Social-Semiotics-An-Introductory-Textbook/Leeuwen /p/book/9780415249447

Vo, H. C. (2025). Resonance in expressions of facial affect and voice affect in "Frozen" animation. *VNU Journal of Foreign Studies*, *41*(1S (Special Issue)), 29-43. https://doi.org/10.63023/2525-2445/jfs.ulis.5390

Zhang, F. (2022). Meaning construction of multimodal synergy in documentary discourse: Taking *The lockdown: One month in Wuhan* as an example. *International Journal of Linguistics, Literature and Translation*, *5*(6). 52-60. https://doi.org/10.32996/ijllt.2022.5.6.7

Webpages

Nguyen Quy (2019). *Netflix series on Asian street food focuses on Saigon*. Retrieved from https://e.vnexpress.net/news/travel/food-recipes/netflix-series-on-asian-street-food-focuses-on-saigon-3908602.html

Tam Anh (2025). *TasteAtlas ranks Vietnamese street foods among the best in Southeast Asia*. Retrieved from https://e.vnexpress.net/photo/food-recipes/tasteatlas-ranks-vietnamese-street-foods-among-the-best-in-southeast-asia-4856205.html

VnExpress (2016). *Saigon's banh mi hailed among the kings and queens of street foods*. Retrieved from https://e.vnexpress.net/news/travel-life/saigon-s-banh-mi-hailed-among-the-kings-and-queens-of-street-foods-3498068.html