Journal of Linguistics and Communication Studies
ISSN 2958-0412
www.paradigmpress.org/JLCS
Volume 5 Number 1 March 2026

The Construction of Virtual Intimacy: A Multimodal Discourse Analysis on the Interactive Mechanism in Virtual Livestreaming

Yuxuan Liu[1]

[1] School of Foreign Studies, University of Science and Technology Beijing, Beijing, China
Correspondence: Yuxuan Liu, School of Foreign Studies, University of Science and Technology Beijing, Beijing, China.

doi:10.63593/JLCS.2026.03.04

Abstract

The emerging phenomenon of virtual livestreaming has garnered academic attention. However, there remains limited multimodal analysis addressing its interactive mechanisms. Grounded in the interpersonal metafunction of systemic functional linguistics and the interactive meaning framework of visual grammar, the present study employs a corpus-based approach to investigate how virtual streamers on YouTube construct a sense of virtual intimacy through verbal, visual, and their integrated multimodal resources. Results show that: (1) verbally, virtual streamers primarily use declarative sentences, medium-value modality, and probability to demonstrate a highly formulaic pattern of emotional expression; (2) visually, they foster an immersive, face-to-face-like interactive atmosphere through frontal, eye-level close-up shots combined with frequent level demands and smiles; (3) correlation analysis reveals a positive correlation between the offer contact and declarative sentence, also between level demand and low-value modality, but a negative correlation is found between smiles and declarative sentences, downward demands and interrogative sentences; (4) the multimodal coordination enables virtual streamers to effectively transform emotional interactions into sustained consumption engagement with virtual symbols. This study offers a novel perspective for research on the virtual streaming industry and digital consumption culture.

Keywords: virtual streamers, multimodal, interactive mechanism

1. Introduction

Driven by the convergence of ACG (anime, comics, and games) subculture, livestreaming culture, and online tipping practices, virtual streamers have rapidly emerged as a new form of cultural consumption. They are typically operated by human performers, with computer-generated animated avatars representing their on-screen presence, and interact with audiences via livestreaming platforms (Qin, 2022). Their revenue streams usually come from audience tipping, platform commissions, sales of related merchandise, etc. This novel form of virtual entertainment can provide the audience with a new interactive and consumption experience while eliminating the physical constraints of traditional streamers.

The origins of the virtual streaming industry can

be traced back to Japan. The performers were initially active on YouTube so they became widely known as "Vtubers". Distinct from traditional virtual idols produced through audio database synthesis, virtual streamers rely on real human performers (called *Nakanohito*) who animate their avatars in real time through motion capture and voice acting technology. This way allows for more realistic and dynamic livestreamed performances. Their virtual appearances (called "avatars") are typically designed by production companies, which can be two-dimensional, three-dimensional, or even animal forms with human characteristics, such as the popular "shark girl" avatar (Tan & Greene, 2025). This diversified character design enhances the audience's visual experience while offering performers greater creative space. It can be seen that the rapid growth of this industry is largely attributed to its unique interactivity and high levels of audience engagement. The audience is attracted by the creative content, interactivity, and the aesthetic appeal of virtual avatars (Peng & Chen, 2024). Besides, the anonymity afforded by these virtual avatars grants performers greater freedom of expression, enabling them to establish a close relationship with audiences without being limited by real-world identity.

In addition to attracting a large number of viewers, the rapid development of the virtual livestreaming industry has also drawn increasing scholarly attention. Existing research mainly centers on the characteristics of virtual streamers and the mechanisms of their interaction with the audience. Studies on their characteristics have examined how factors such as attractiveness (Liu & Zhao, 2025), coolness (Gao et al., 2025), emotionally expressive language styles (Gong & Sun, 2025), and anthropomorphism (Chen et al., 2024; Peng & Chen, 2024) influence the audience's consumption behavior. For instance, Liu and Zhao (2025) measured attractiveness of virtual streamers through dimensions of similarity, familiarity, and likability, as well as the two complementary aspects of voice and speech, showing that these factors can positively affect the audience's tipping behavior and consumption intentions. Moreover, higher textual similarity and familiarity can enhance the sense of closeness between virtual streamers and their audience, thus playing a potential role in prompting viewers' initiative to give financial support. Through a series of experimental studies, Gong and Sun (2025) found that the emotional language used by virtual streamers leads to a higher consumer intention to follow the advice (CIFA). Emotional language, by fostering emotional resonance, is considered more effective in stimulating audiences' purchase intentions than rational language. It further promotes consumption behavior by enhancing the audience's perceived agency and perceived experience. What's more, virtual streamers, due to their unique motion-capture technology, can create a hyperreal experience through highly realistic visual avatars and actions, thus strengthening viewers' immersion and making it easier for them to accept recommended products and services (Chen et al., 2024; Peng & Chen, 2024). However, the semiotic production process of these streamers has also brought an array of critical issues. For example, the subjectivity of the "person behind the avatar" (*Nakanohito*) is often erased, with their real physical body hidden. As a result, the virtual persona consumed by the audience has actually become a highly patterned digital body (Peng & Chen, 2024).

On the other hand, scholars have also explored the interaction mechanisms between virtual streamers and their audience. The following three key factors play a crucial role in the establishment of the interaction process. First, the symbolic interaction motivated by virtual bodies. From the perspective of communication semiotics, Qin (2022) points out that virtual streamers, based on the inherent symbolic characteristics of their virtual bodies, enable their audience to achieve a certain degree of cultural identity through self-projection. The audience, as the recipient of symbols, may also psychologically separate themselves from the real world and invest considerable emotional efforts to maintain the virtual world jointly established by virtual streamers and their fan groups (Peng & Chen, 2024). Second, the live content itself can enhance audience participation (Chen et al., 2025; Lu et al., 2021). Based on Consistency Theory and Dramaturgical Theory, Chen et al. (2025) found that the consistency between the viewer's interest-live content congruence (IC) and viewer's value-streamer's value congruence (VE) contributes to the audience's immersion, which in turn affects their attitude and behavior intentions. Third, the role-playing ability and performance of virtual streamers can also facilitate interaction with the audience. Their ability to convincingly present specific persona settings during livestreaming

tends to positively moderate the relationship between IC and immersion, while negatively moderating the relationship between streamer's persona-live content congruence (PC) and immersion (Chen et al., 2025). The result suggests that the role-playing ability of virtual streamers can bolster audience involvement in interaction. In addition, they are better able to capture viewers' attention and evoke emotional resonance through the performance of their virtual personas (Lu et al., 2021). To sum up, existing research on virtual streamers has showed their interest in individual characteristics and the single dimension of interaction with the audience. Yet, these studies tend to overlook the importance of multiple semiotic resources collaboratively constructing meaning. Given that these streamers' interaction mechanisms involve the coordination of multiple modalities such as language, images and embodied performance, how these modalities work together to enhance the intimacy between virtual streamers and their audience remains an area that needs further exploration.

Modality refers to the channels and media of communication, including systems of signs such as language, technology, images, color, and music (Zhu, 2007). Multimodality, in turn, concerns how meaning is constructed through the interplay of various semiotic resources, which is the use of multiple modes of communication in the design of symbolic artifacts or events and how these modes are combined in specific, structured ways (Kress & van Leeuwen, 2006). Current research on multimodality mainly draws from three different perspectives. First of all, the social semiotic approach advocated by Kress and van Leeuwen (2006) emphasizes how different modalities interact to produce complex meanings based on systemic functional linguistics. This framework attends to how social and cultural factors shape the use of semiotic resources and provides a systematic approach for analyzing the design of multimodal combinations. For instance, the "visual grammar" they proposed provides a structured framework for interpreting how visual elements convey meaning. The second perspective, derived from O'Toole (1994), applies the systemic functional grammar framework to visual arts by breaking down works like paintings into hierarchical compositional levels. This was later extended by O'Halloran (2005) to describe the grammatical systems of various semiotic resources and their metafunctions. The third approach is Norris's (2004) multimodal interaction analysis. It builds on interactional sociolinguistics and mediated discourse analysis to investigate the collaborative effects of language, gesture, movement, and other modalities in human communication, with a particular focus on how meaning is co-constructed in natural interactions.

To date, the scope of multimodal discourse analysis has expanded beyond static artifacts such as picture books (Qi, 2022), comics (Zhao, 2022), literary texts (Gu & Catalano, 2022), image-text advertisements (Kenalemang-Palm, 2023), and AI-generated images (Putland et al., 2025), to increasingly include dynamic audiovisual media, especially live streaming. For example, Wang and Pan (2022) employed conversation analysis to demonstrate that the multimodal linguistic interaction in e-commerce live streaming constitutes an effective persuasive strategy. Through frequent use of interactive symbols and emotionally personal expressions, streamers guide audiences toward purchasing behaviors, turning linguistic interaction into a form of economic vitality. Huang et al. (2020) also found that the success of Chinese livestreamer Li Jiaqi illustrates how gender identity can be mobilized as a resource to attract audiences, challenging conventional gender norms and expectations and offering new directions for the development of livestream e-commerce. Besides, in constructing the interactive significance of e-commerce livestreams, verbal modality serves to provide information and demonstrate objectivity, while visual modality functions to capture consumer attention and reduce interpersonal distance (Sun, 2024). However, it is vital to note that existing multimodal analyses of livestreaming interactions have largely focused on human streamers.

Based on previous studies, it is apparent that current research gaps exist in two main areas. On the one hand, research on virtual streamers has mainly converged on macro-level, singular characteristics such as coolness and attractiveness, or examined audience interaction mechanisms from a single perspective. But the collaborative effects of multimodal features such as language, facial expressions, and other resources remain underexplored. On the other hand, the application of multimodal discourse analysis within the livestreaming context has largely centered on human streamers, with comparatively little attention paid to virtual

streamers. In response to these gaps, the present study adopts the frameworks of interpersonal metafunction from systemic functional linguistics and interactive meaning from visual grammar to investigate how the multimodal features of verbal and visual modalities, as well as their interplay, construct the intimacy between virtual streamers and their audience. This study not only helps to broaden the research perspectives on virtual streamers but also offers more practical insights for the livestreaming industry and digital consumer culture. The following research questions will be addressed:

RQ1: What are the characteristics of verbal and visual modalities in virtual livestreaming?

RQ2: How do virtual streamers jointly construct a sense of intimacy with their audience through the collaboration of verbal and visual modalities?

2. Data and Methods

2.1 Data Description

Although virtual streamers currently broadcast on platforms such as YouTube, Twitch, and Bilibili, Vtubers hold a dominant position within the livestreaming market. According to data, Vtubers account for 80% of the top 10 streamers (Playboard, 2022b). In view of this, the present study draws its corpus from the real-time livestreams of English-speaking virtual streamers on the YouTube platform. Considering the wide range of livestreaming genres, including performances, chats, and games, etc., and their differences in interaction density, this study mainly focuses on English chatting livestreams, which feature a high proportion of verbal communication and frequent interaction. These characteristics make them particularly suitable for examining the construction of virtual intimacy.

In order to capture the interactive characteristics of streamers with different identity attributes, a total of five livestream videos were selected from Nijisanji EN's company-affiliated Vtubers and independent Vtubers unaffiliated with any management company. During data collection, about 10 minutes of high-interaction clips were randomly extracted from each video to ensure that the clips contain more concentrated audience comments and Super Chat (SC) triggers. After collecting data, this study built a multimodal corpus, which included video and text data. The total duration of video data is 3,073.488 seconds, and the text data totals 257,710 tokens, covering the streamers' spoken English output.

To ensure the data quality and the validity of subsequent analysis, the following steps were carried out in the corpus preprocessing. First, on the basis of YouTube's auto-caption, the subtitle text was transcribed through manual proofreading, with sentence-by-sentence comparison against the original video audio to correct errors related to tone words, ellipses, automatic recognition errors, etc. Next, the obtained text data were cleaned by removing meaningless filler words (e.g., *uh*, *hmm*), redundant characters (e.g., *wwww*), and background noise. Then, for multimodal interactive meaning analysis, video clips were converted into frame-by-frame visualization formats, and a video annotation table was created and synchronized with the corresponding textual timeline. The final multimodal corpus thus covers verbal, visual, and verbal-visual collaborative modal features.

2.2 Analytical Framework

2.2.1 Interpersonal Metafunction in Systemic Functional Linguistics

The current study integrates two multimodal analytical approaches, social semiotics and systemic functional grammar, and adopts the frameworks of interpersonal metafunction in systemic functional linguistics and interactive meaning in visual grammar as its analytical basis. Systemic functional linguistics (SFL), proposed by Halliday, emphasizes the social semiotic functions of language and categorizes its metafunctions into three types: ideational, interpersonal, and textual. Among these, the interpersonal metafunction focuses on how language constructs relationships between speakers and listeners, expresses attitudes, and negotiates interactional strategies in communicative processes. Based on Halliday (2004), this study analyzes the verbal modality of virtual livestreaming discourse by examining two subsystems in the interpersonal meaning framework: mood and modality.

The mood system reflects the fundamental interactive structure between speakers and listeners, indicating the type of speech function being enacted in communication. It primarily involves three types: declaratives, interrogatives, and imperatives. Declaratives are used to convey information or express opinions, interrogatives serve to elicit responses or prompt interaction, while imperatives typically issue requests or commands. The modality system, on the other

hand, encodes the speaker's subjective attitude and degree of commitment toward the proposition, expressing meanings that fall between affirmation and negation. Halliday (2004) divides modality into three levels: high, median, and low, which correspond to the speaker's strong, moderate, and weak attitudes towards probability, usuality, obligation, or inclination. Among them, probability and usuality belong to modalization, while obligation and inclination fall under modulation. For example, modal adverbs such as *definitely*, *probably*, and *maybe* or modal verbs like *can*, *must*, and *should* can convey the speaker's varying degrees of judgment towards the topic. Grounded in the interpersonal meaning system of SFL, this study selects mood and modality as core analytical indicators to examine the interpersonal interaction strategies in the verbal modality of virtual livestreaming and to explore how these verbal forms affect the interactive order and atmosphere in the virtual environment.

2.2.2 Interactive Meaning in Visual Grammar

Drawing on Halliday's theory of metafunctions, Kress and van Leeuwen (2006) proposed the visual grammar framework, which systematically explicates how images realize representational, interactive, and compositional meanings. This study focuses on the interactive meaning system within this framework, which include three key elements: contact, social distance, and attitude. These elements are employed to analyze the interactive strategies used by virtual streamers through visual resources to construct a simulated sense of intimacy during livestreams.

Contact refers to whether the character in the image engages in eye contact with the audience. Kress and van Leeuwen categorize this into two types: *demand* and *offer*. In *demand* images, the character directly gazes at the viewer, making an emotional or cognitive demand to enhance interactivity. By contrast, *offer* images position viewers as observers, as the character does not look directly at them. Social distance determined by shot scale is divided into three categories: *close-up*, *medium shot*, and *long shot*. Close-up shots convey a strong sense of intimacy by enlarging to the face or shoulders; medium shots present the part above the knees to create a moderate social distance; and long shots capturing the full body and surrounding environment tend to weaken closeness. Attitude concerns the angle of image shooting, which implies the participant's stance, including *horizontal angle* and *vertical angle*. The horizontal angle includes *frontal* and *oblique* angles, while the vertical angle involves *high*, *eye-level*, and *low* angles.

According to the original framework, it is worth noting that the vertical angle refers to the position of the camera rather than changes in eye gaze direction. However, given virtual streamers are limited by motion-capture technology in virtual livestreaming, camera angles remain basically fixed. As such, this study refines the original classification of contact and further divides *demand* into three categories based on gaze direction: *level demand*, *upward demand*, and *downward demand*. Besides, this study also supplements facial expressions like smiles and surprises as indicators of affective engagement. The detailed analytical framework is presented in Table 1.

Table 1. Analytical framework of the present study

Modality type	Dimension	Subcategory	Source
Verbal modality	Mood	declarative, interrogative, imperative	Halliday (2004)
	Modality	high, medium, low	
Visual modality	Contact	demand (level angle, downward angle, upward angle), offer	Kress and van Leeuwen (2006), self-defined
	Social distance	close shot, medium shot, long shot	
	Attitude	horizontal angle, vertical angle	
	Facial expression	smile, surprise, etc.	Self-defined

2.3 Procedure

Grounded in the multimodal discourse analysis framework above, this study conducts a systematic analysis of the collected corpus. First, the annotation process was carried out on the self-built multimodal corpus, covering both verbal and visual modalities. For the verbal modality, mood and modality features in the Vtubers' spoken discourse were identified by combining the UAM Corpus Tool 6 and manual annotation, following the interpersonal metafunction framework of systemic functional linguistics. Mood is classified into three types, declarative, interrogative, and imperative, while modality is categorized into high, medium, and low levels according to Halliday's interpersonal system.

For the visual modality, the study annotated features of contact, social distance, attitude, and facial expressions through the ELAN 6.7 multimodal annotation tool based on the interactional meaning system in Kress & van Leeuwen's (2006) visual grammar framework. In terms of contact, this study refined the original classification and divided *demand* into *level demand*, *downward demand*, and *upward demand* to capture whether streamers make direct eye contact with the audience and the direction of the gaze. Social distance is categorized according to shot scale into *close-up*, *medium shot*, and *long shot*. Attitude includes both *horizontal angle* and *vertical angle*. The facial expression category covers smiles, surprises and other common emotional expressions. All annotations align frame by frame with the corresponding video and text data to ensure precise correspondence between the verbal and visual modalities. To guarantee reliability, each annotation result was reviewed and proofread three times.

After annotation, the frequency and proportion of each verbal and visual feature were calculated through the data analysis function of UAM Corpus Tool 6 and ELAN 6.7 respectively. Finally, a Spearman correlation analysis was conducted in SPSS 27 to examine the relationships between features across the two modalities. The study also generated the heatmap via Python to reveal the distribution patterns and collaborative mechanisms of multimodal features in virtual livestreaming.

3. Results

3.1 Distribution of Verbal Features

Table 2. Distribution of mood system

Mood Clauses	N	%
Declarative clause	942	88.42
Interrogative clause	60	3.89
Imperative clause	82	7.69
Total	1080	100

This study conducts frequency statistics on the verbal modality features in virtual livestreams. In terms of the mood system, as seen in Table 2, virtual streamers primarily use declarative sentences (88.42%) during their livestream interactions. This result suggests that virtual streamers tend to communicate with their audience through direct and explicit statements. By contrast, imperative sentences account for 7.69%, while interrogative sentences are the least used, accounting for only 3.89%. This distribution feature may reflect virtual streamers' preference for maintaining the rhythm of conversations and the dominance of topics through declarative language.

Table 3. Distribution of modality system

Values of Modality	Probability	Usuality	Obligation	Inclination	Total	%
High	32	19	9	23	83	40.49
Medium	60	1	1	44	106	51.71
Low	14	2	0	0	16	7.80

Total	106	22	10	67	205	100
%	51.70	10.73	4.89	32.68	100	

Regarding the modality system, the data in Table 3 indicate that the use frequency of medium-value modality is the highest (51.71%), followed by high-value modality (40.49%), while low-value modality appears the least (7.80%). In addition, among the four modality types, probability is employed most often (51.70%). For instance, the highly frequent use of expressions of medium to high modality such as *I'm sure* and *probably* suggests that virtual streamers tend to avoid absolute or overly assertive statements during livestreams. In contrast, obligation modality is used the least (4.89%), which indicates that virtual streamers generally refrain from making compulsory requirements for their audience. In a nutshell, the verbal modality in virtual livestream interactions is characterized by a highly declarative discourse style, with a preference for medium to high-value modality expressions that carry a degree of uncertainty.

3.2 Distribution of Visual Features

Table 4. Distribution of visual features

Interactive Meanings	**Realizations**	**N**	**Time (s)**	**%**
Contact	level demand	159	2069.217	47.60
	downward demand	31	180.754	9.28
	upward demand	5	10.685	1.50
	offer	139	812.832	41.62
Social distance	close shot	5	3073.488	100
Horizontal angle	frontal angle	5	3073.488	100
Vertical angle	eye-level angle	5	3073.488	100
Facial expression	smile	105	235.793	87.50
	surprise	15	32.572	12.50

According to Table 4, the interaction of virtual streamers at the visual level is presented in a fixed close-up shot, frontal angle, and eye-level angle. In most cases, the streamer occupies the central position on screen and looks directly at the audience. The analysis of contact shows that virtual streamers primarily interact with viewers through demand gaze, with level demand accounting for the highest proportion (47.60%) and the longest duration (2069.217). This indicates that virtual streamers mainly simulate eye contact with the audience through direct, level gazes. But it is found that the proportion of downward demand (9.28%) and upward demand (1.50%) is obviously low, suggesting that their visual interaction strategies tend to emphasize equality in social relations. For facial expressions, smiles dominate at 87.50%, which constitutes the main emotional expression for virtual streamers, while expressions such as surprises (12.50%) are typically used to enhance the performative and dramatic aspects of the livestream through dynamic facial changes.

3.3 Spearman Correlation Between Verbal and Visual Features

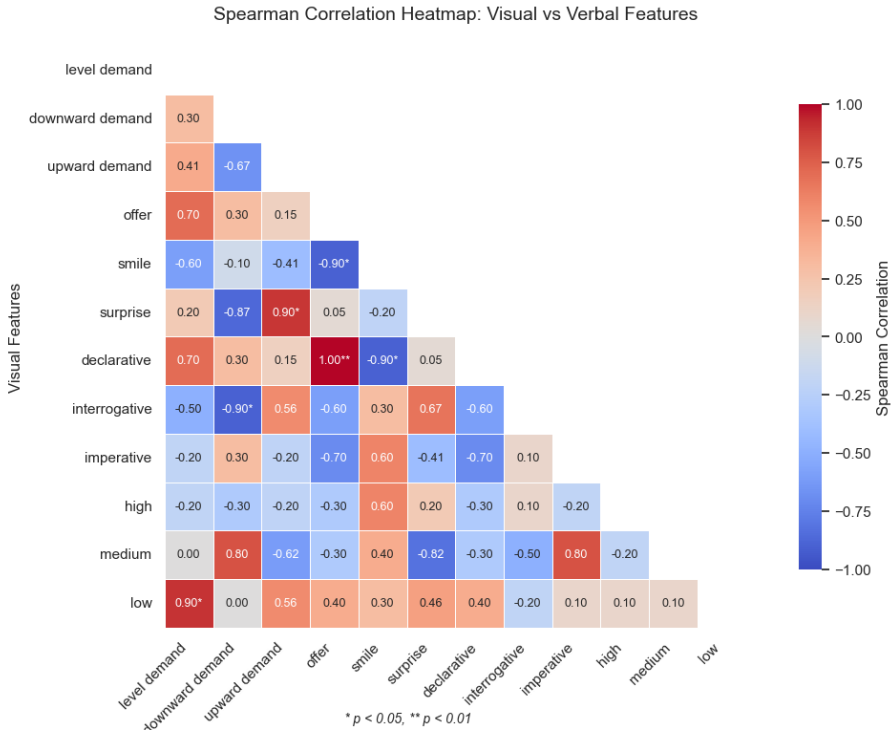

Figure 1. The heatmap for spearman correlations between verbal and visual features

In the analysis of the relationship between the verbal and visual layers, this study employs Spearman's rank correlation coefficient test to examine the correlations between multimodal features in virtual livestreams. As shown in Figure 1, several significant correlations are identified between verbal and visual features. There is a strong positive correlation between the streamer's offer gaze and the use of declarative sentences (r = 1.000, p = 0.000), indicating that virtual streamers often cooperate with an averted gaze to create a relaxed and casual interactive atmosphere when making information statements. In addition, a significant positive correlation is observed between level demand and low-value modality expressions (r = 0.900, p = 0.037). In other words, virtual streamers tend to maintain a level gaze when expressing uncertainty or adopting a more tentative attitude.

On the contrary, smiles show a significant negative correlation with declarative sentence structures (r = -0.900, p = 0.037). This implies that virtual streamers may smile less at the same time when making declarative statements possibly for the sake of maintaining seriousness or authority while delivering information. Similarly, downward demand is also significantly negatively correlated with the use of interrogative sentences (r = -0.900, p = 0.037), indicating that virtual streamers rarely use a downward gaze when posing questions to the audience, likely in order to avoid a sense of dominance. These findings reveal the coordinated distribution features of verbal and visual modalities in virtual livestreams.

4. Discussion

4.1 Constructing Intimacy Through Language: Mood and Modality in VTubers' Performances

The intimacy constructed by virtual streamers through the verbal modality centers on the strategic coordination of the mood system and modality system. From the perspective of systemic functional linguistics, the verbal features of virtual streamers exhibit the coexistence of highly formulaic and emotional patterns, and this contradictory unity forms the very foundation of their sense of intimacy.

In terms of the mood system, virtual streamers predominantly employ declarative sentences, which is consistent with Shi (2024)'s findings. These declaratives not only serve to transmit information, express opinions, and respond to live comments but also act as carriers of emotional projection. For example, in this study's corpus, expressions like *I remember I read the first two books* and *I want to go back to university* focus on the streamer's personal experiences and subjective feelings, avoiding the language pressure that imperative statements might

impose on the audience. It is evidenced that streamers have established an interactional framework centering on sharing rather than informing. Such intimate narrative behavior invites the audience into streamers' emotional world through self-sharing to bring the psychological distance closer to each other. It also indicates that they attempt to use declarative sentences to control the rhythm of livestreams and refocus attention on themselves in a timely manner (Qin, 2022). It is worth noting that many declaratives can be regarded as stylized templates of intimate speech even if they seem to be personal habitual expressions. For instance, virtual streamers often repeatedly use *I'm so proud of you*, *I'm really proud of you*, and *I'm glad that you enjoy hearing my voice* when responding to Super Chats. These expressions evidently form a standardized model of one-way emotional output through the fixed collocation of the subject *I* with affective adjectives like *proud* or *glad*. Such declaratives are frequently accompanied by the accumulation of adverbs of degree (e.g., *so*, *really*, *absolutely*), reflecting how streamers compensate for the affective limitations of their virtual personas in emotional communication via quantitative marking of language intensity.

The use of the modality system in virtual livestreams also displays clear intimacy-oriented features. Medium-value modality appears most frequently, with probability being the most common. This type of modality is often used to convey speculation, assumptions, wishes and other subjective judgments, which help avoid absolute assertions and create space for the audience to agree, supplement, or refute. It can also mitigate the potential hierarchy of discursive power between the streamer and audience. For instance, when discussing the relationship between game IPs, a streamer's vague statement like *most likely that it is a spiritual successor* conveys professional knowledge while avoiding the risks of assertive claims. Virtual streamers also frequently use expressions such as *I think* and *I figure* as typical markers of probability that attribute the language responsibility to personal experience, so as to avoid invasive judgments on the audience's own knowledge. This modality construction mechanism can effectively maintain the equal dialogue relationship and encourage the audience's response and sympathy based on shared perspectives, thereby realizing the dynamic construction of interactive intimacy.

Besides, the most innovative feature of virtual streamer discourse lies in the nested structure of modality expressions. In response to wishes about a medical exam, for example, the streamer built a complex modality chain: *You're going to pass the exam, or you are going to be diagnosed as completely healthy. And if you are not, you will heal and everything will be fine*. This turn of speech contains four consecutive expressions of modality, forming an emotional reinforcement network that ensures the speech itself possesses emotional effectiveness regardless of the actual result. This confirms the view of Gong and Sun (2025) that the use of affective language can enhance the audience's emotional resonance.

4.2 Creating Equality Through Visual Modality: Visual Affection in Virtual Livestreaming

The visual presentation of virtual livestreaming is highly standardized and patterned. It mainly realizes continuous visual engagement with the audience and the imitation of the interactive situation through the use of fixed close-up shots, frontal angles, and eye-level angles, which conforms to the results of Shi (2024). Such visual design effectively dispels the sense of distance between the streamer's virtual identity and real-time interaction, offering viewers an immersive experience of face-to-face-like communication during watching.

Figure 2. The screenshot of the VTuber's livestreaming

The consistent use of fixed close-up shots, frontal angles, and eye-level angles ensures that facial expressions and gaze direction of virtual streamers always occupy the center of the picture through deliberately imitating the perspective of video calls, which visually frames the avatar within the visual range of intimate distance. In practice, the streamer image usually occupies about 60-70% of the area in the center of the picture and the background matches the avatar's image setting (see Figure 2). Compared with traditional livestreams, this visual arrangement not only enhances the subjectivity of streamers but also serves a prominent emotional focus function that can amplify micro-expressions and emotional changes for the audience to detect during the interaction. However, considering the realistic technical factors of virtual livestreaming, such a fixed visual design is closely related to the actual space of the human performer in front of the desktop computer. Due to the limitations of motion capture technology, it is rare for virtual streamers to fully display whole-body movements or the interactions with the physical world (Lu et al., 2021).

In livestreaming, the demand gaze constitutes the most fundamental visual pattern in virtual streamer interactions, with level demand accounting for as high as 47.60% and sustaining the longest duration. The level demand is commonly associated with equality, friendliness, and closeness. Virtual streamers' eye movements can be controlled by the capture device to look directly at the audience in a straight-facing way, forming an "eye-to-eye" interaction. Moreover, in interactive moments such as replying bullet comments, the system often triggers specific eye animations, such as the effect of pupil dilation or blinking, to simulate nonverbal reactions when receiving positive feedback. Through this eye contact with the audience, virtual streamers construct a silent yet emotional interaction atmosphere, which makes viewers feel the emotional value of "being looked at" and "being listened to". Therefore, viewers' psychological connection and trust in the streamer can be established while also dissociating the implicit hints of power.

Figure 3. The screenshot of a smiling VTuber

In addition, smile is the most frequently employed facial expression in virtual livestreams. As a nonverbal symbol generally interpreted in a positive way, smile serves to alleviate social tension and foster intimacy. Virtual streamers use continuous smiles to convey a friendly and welcoming attitude to viewers. The high frequency of smiles is likely to elevate the emotional atmosphere and also effectively mitigates the interaction barriers inherent in virtual identity, so that viewers are allowed to experience simulated real-time conversations. For example, when receiving rewards or comments from the audience, virtual streamers are typically accompanied by smiling expressions and a frontal and eye-level angle, which can instantly narrow the emotional distance between themselves and the audience (see Figure 3). However, as Lu et al. (2021) point out, virtual livestreams is considered a hybrid form of integration of the virtual and real world, where technical glitches such as avatar clipping or persistent smiling due to model design may result in moments where the avatar appears to be smiling, but the performer may in fact be looking down, or the avatar might be designed to maintain a constant smile to perpetuate a sense of warmth and approachability.

4.3 Multimodal Coordination: Virtual Intimacy and Affective Consumer Mobilization

Through the results of the Spearman correlation analysis, it can be found that there is a significant coordination pattern between the language

selection and visual presentation of virtual streamers. And the interactive atmosphere created by different combination strategies can directly affect the audience's emotional experience and perceptions of intimacy.

First of all, a strong positive correlation is found between the virtual streamer's offer gaze and the use of declarative sentences. This indicates that when virtual streamers avert their gaze from direct eye contact, they are more inclined to use declarative sentences to share personal experiences, express opinions, or respond to comments. This combination strategy reduces the pressure and directness of speech, which makes the interaction appear more casual to facilitate a relaxed and intimate atmosphere. For instance, in terms of some scenes like where the streamer say *I wanted to watch cartoons with a plot with excitement you know what I mean*, the casual words and non-direct gaze place the audience within an informal and friendly communication environment where intimacy is more readily generated.

There is also a significant positive correlation observed between level demand and low-value modality. When engaging in direct level gaze with viewers, virtual streamers are customary to use expressions implying possibility or speculation which lower the compulsiveness and evaluation of language. This strategy weakening verbal pressure may further promote the audience's willingness to participate in the interactive process, and even stimulate tipping behavior (Qin, 2022). It is crucial that this verbal-visual coordination is especially evident when Super Chats or tipping notifications appear. Whenever a viewer sends a tip or gift and the system prompt appear, virtual streamers often immediately switch to a level demand, combine it with a smile, and use highly emotive thank-you words and nicknames, such as *thank you so much for the five gifted membership thank you so much thank you*. Although the modality value remains fixed, the intensity of emotional expression is dramatically enhanced, the smile is emphasized, and the gaze shifts instantly from offer to level demand. This instantaneous multimodal linkage produces an effect of "you are paid special attention", which effectively activates viewers' emotional identification mechanism. This mimic intimacy assumes an important consumption-driven function under the commercial logic of virtual livestreaming, that is, the audience is prone to continuous consumption behavior under the induction of virtual closeness. This suggests that the multimodal intimacy strategy employed by virtual streamers function not only as tools for emotional interaction but also as integral components of the consumption logic supported by emotional labor in the digital livestream economy. In this process, viewers seem to be autonomous consumers, yet in fact they remain entranced by the illusion of subjectivity constructed through images (Yang, 2011).

In contrast, there exists a significant negative correlation between smiles and declarative sentences. It means that when virtual streamers smile, they are more likely to cooperate more with highly interactive utterances such as questions, exclamations, or emotional calls to mobilize the audience's response and maintain a lively interaction rhythm. For example, a smiling streamer asks, "what did you do and what do you want to do if you have a time machine to go back to high school life," which motivates the audience to participate through real-time facial expressions and interactive speech. While strengthening intimacy, it also reinforces the audience's sense of identity as a consumer subject in the consumption-oriented context. What's more, downward demand negatively correlates with the use of interrogative sentences because a downward gaze conveys authority and distance, which is not conducive to stimulating viewers' response. Yet, as highly interactive utterances, interrogative sentences need to rely on a relaxed and equal communication environment. In brief, by leveraging the multimodal intimacy mechanism, virtual streamers shape their personas into objects that can be emotionally projected and psychologically identified by the audience, and continue to stabilize the subject-object structure of the livestreaming economy.

5. Conclusion

Based on the interpersonal metafunction of systemic functional linguistics and the interactive meaning in visual grammar, this study using a corpus-based approach conducts a multimodal discourse analysis of virtual livestreams to explore how virtual streamers construct intimacy through the verbal, visual modality, and the coordinated interplay of both. The main findings are as follows: (1) at the verbal level, virtual streamers foster a friendly and equal interactive atmosphere by frequently using declarative sentences, medium-value modality, and probability mood speech; (2) at the visual level, streamers consistently employ frontal, eye-level

close-up shots, level demand and smiles, effectively creating an intimate experience like face-to-face communication; (3) at the level of multimodal coordination, there is a strong positive correlation between offer gaze and declarative sentences, and between level demand and low-value modality; in contrast, a significant negative correlation is found between smiling expressions and declarative sentences, also between downward demand and interrogative sentences; (4) virtual streamers release emotional intimacy signals through multimodal means, which bolsters the audience's awareness participation and implicitly mobilizes emotional engagement to encourage continuous consumption. It thus becomes evident that intimacy in virtual livestreaming is not merely a form of emotional interaction, but a kind of simulated emotional labor accompanied with consumption mobilization.

This study, however, has certain limitations. On the one hand, the sample data mainly derives from English-speaking virtual livestreams. Cross-cultural differences in language and norms may affect the universality of multimodal interaction strategies, which can be extended to the livestream content of virtual streamers in different language and cultural contexts. On the other hand, the size of samples in this study is relatively limited, which may result in insufficient data coverage. However, this study enriches the understanding of the multimodal interaction mechanisms in the virtual livestreaming industry, and also provides a new perspective for the phenomenon of emotional economy in digital consumption culture.

References

Chen, H., Shao, B., Yang, X., Kang, W., & Fan, W. (2024). Avatars in live streaming commerce: The influence of anthropomorphism on consumers' willingness to accept virtual live streamers. *Computers in Human Behavior*, *156*, 108216. https://doi.org/10.1016/j.chb.2024.108216.

Chen, Y., Li, L., & Zhou, W. (2025). Impact of viewer-streamer-content congruence on users' behavioral intention in virtual streaming: The moderating effect of role-playing. *Electronic Commerce Research and Applications*, *70*, 101492. https://doi.org/10.1016/j.elerap.2025.101492.

Gao, W., Jiang, N., & Guo, Q. (2025). How cool virtual streamer influences customer in live-streaming commerce? An explanation of stereotype content model. *Journal of Retailing and Consumer Services*, *82*, 104139. https://doi.org/10.1016/j.jretconser.2024.104139.

Gong, X., & Sun, P. (2025). Can virtual streamers express emotions? Understanding the language style of virtual streamers in livestreaming e-commerce. *Journal of Retailing and Consumer Services*, *82*, 104148. https://doi.org/10.1016/j.jretconser.2024.104148.

Gu, X., & Catalano, T. (2022). Representing transition experiences: A multimodal critical discourse analysis of young immigrants in children's literature. *Linguistics and Education*, *71*, 101083. https://doi.org/10.1016/ j.linged.2022.101083.

Halliday, M. A. K. (2004). *An introduction to functional grammar*. Routledge.

Huang, H., Blommaert, J., & Van Praet, E. (2020). "OH MY GOD! BUY IT!" a multimodal discourse analysis of the discursive strategies used by Chinese ecommerce live-streamer Austin Li. In C. Stephanidis, G. Salvendy, J. Wei, S. Yamamoto, H. Mori, G. Meiselwitz, F. F.-H. Nah, & K. Siau (Eds.), *HCI International 2020 – Late Breaking Papers: Interaction, Knowledge and Social Media* (pp. 305–327). Springer International Publishing.

Kenalemang-Palm, L. M. (2023). The beautification of men within skincare advertisements: A multimodal critical discourse analysis. *Journal of Aging Studies*, *66*, 101153. https://doi.org/10.1016/j.jaging.2023.101153.

Kress, G., and van Leeuwen, T. (2006). *Reading images: The grammar of visual design*. (2nd ed). Routledge.

Liu, H., & Zhao, J. (2025). VTuber attractiveness and its effect on viewer gifting. *International Journal of Human–Computer Interaction*, 1–16. https://doi.org/ 10.1080/10447318.2025.2465866.

Lu, Z., Shen, C., Li, J., Shen, H., & Wigdor, D. (2021). More kawaii than a real-person live streamer: Understanding how the otaku community engages with and perceives virtual YouTubers. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/ 3411764.3445660.

Norris, S. (2004). *Analyzing multimodal interaction: A methodological framework*. Routledge.

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

O'Toole, M. (1994). *The language of displayed art*. Leicester University Press.

Peng, J., & Chen, J. (2024). Research on the body aesthetics of virtual hosts and the mechanisms of symbolic consumption: A Baudrillardian perspective. *Journal of Southwest Minzu University (Humanities and Social Sciences Edition)*, *45*(11), 144–153. https://kns.cnki.net/KCMS/detail/detail.aspx?dbcode=CJFQ&dbname=CJFDLAST2025&filename=XNZS202411016.

Playboard. (2022. Dec 16). Most super chatted. *Playboard*. https://playboard.co/en/ youtube-ranking/most-superchatted-all-channels-in-worldwide-total.

Putland, E., Chikodzore-Paterson, C., & Brookes, G. (2025). Artificial intelligence and visual discourse: A multimodal critical discourse analysis of AI-generated images of "dementia". *Social Semiotics*, *35*(2), 228–253. https://doi.org/10.1080/10350330.2023.2290555.

Qi, F. (2022). A study of traditional artistic visual narratives in original picture books based on multimodal discourse analysis. *Publishing Journal*, *30*(3), 43–50. https://doi.org/10.13363/j.publishingjournal.20220517.012.

Qin, Y. (2022). Virtual body and semiotic consumption: A study of interactive mechanisms in virtual streamer livestreams. *Application Research*, *8*(2), 39–44. https://doi.org/10.16604/j.cnki.issn2096-0360.2022.02.006.

Shi, Y. (2024). Research on live broadcast discourse of network virtual uploader from a multimodal perspective. [Unpublished Master's thesis]. Shandong University.

Sun, N. (2024). A multimodal analysis on the interactive meaning of e-commerce live-streaming discourse. [Master's thesis, Harbin Engineering University]. https://doi.org/10.27060/d.cnki.ghbcu.2023.000833.

Tan, Y. H., & Greene, B. R. (2025). Can a 2D shark girl be an influencer? Uncovering prevailing archetypes in the virtual entertainer industry. *Journal of Business Research*, *186*, 114951. https://doi.org/10.1016/j.jbusres.2024.114951.

Wang, Y., & Pan, D. (2022). Research on multimodal interaction in Taobao live streaming. *Chinese Journal of Language Policy and Planning*, *7*(3), 34–46. https://doi.org/10.19689/j.cnki.cn10-1361/h.20220303.

Yang, H. (2011). Hyperreality, simulations and implosion — three key concepts in the late Jean Baudrillard's thought. *Jiangsu Social Sciences*, (4), 14–21. https://doi.org/10.13858/j.cnki.cn32-1312/c.2011.04.019.

Zhao, X. (2022). Ecological discourse analysis based on multimodal metaphor scenarios: a case of bioenergy political cartoons. *Foreign Languages in China*, *19*(6), 60–69. https://doi.org/10.13564/j.cnki.issn.1672-9382.2022.06.003.

Zhu, Y. (2007). Theory and methodology of multimodal discourse analysis. *Foreign Language Research*, *5*, 82–86. https://doi.org/10.16263/j.cnki.23-1071/h.2007.05.034.