

Interactive Mechanisms in Cross-Border E-Commerce Livestreaming on POP MART U.S. SHOP

Xiangyang Zhang

School of International Studies, Soochow University, Suzhou, Jiangsu 215006, China.

*Corresponding Author: Xiangyang Zhang

DOI: <https://doi.org/10.5281/zenodo.23347473>

Article History	Abstract
Original Research Article	<i>E-commerce livestreaming has been widely researched in recent years from the perspectives of platform development, marketing strategy and linguistic features. However, scant attention has been given to how actions are organized in a live stream. To address this gap, this study adopted a qualitative approach to investigate POP MART U.S. SHOP livestreams on TikTok through the lens of multimodal interaction analysis. It systematically examined how multiple communicative modes are coordinated to construct actions across different interactive contexts and found that communicative modes are mobilized in accordance with the communicative purpose of each livestreaming stage. Rather than being realized through a simple accumulation of isolated modes, livestreaming interaction is constructed through the dynamic coordination of multiple lower-level actions, which together shape foregrounded higher-level actions. Moreover, modal density varies across different actions and stages, which is determined not merely by the duration of a single mode but by the intensity and complexity of modal coordination. Based on these findings, this study is expected to extend the application of multimodal interaction analysis to e-commerce and offer practical guidance for Chinese brands to optimize their livestreaming communication strategies in overseas markets.</i> Keywords: Multimodal interaction analysis; higher-level actions; modal density; e-commerce live-streaming; POP MART; TikTok.
Received: 05-08-2026	
Accepted: 15-09-2026	
Published: 13-10-2026	
Copyright © 2026 The Author(s): This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC BY-NC) which permits unrestricted use, distribution, and reproduction in any medium for non-commercial use provided the original author and source are credited. Citation: Xiangyang Zhang. (2026). Interactive mechanisms in cross-border e-commerce livestreaming on POP MART U.S. SHOP. UKR Journal of Economics, Business and Management, 2(10), 103-118.	

1. Introduction

1.1 Research on cross-border e-commerce livestreaming

With the rapid growth of the e-commerce market, livestreaming has become an increasingly influential form of online commercial communication ^[1]. In China, the rapid expansion of livestreaming platforms has intensified competition in the domestic market. Taking Douyin, the Chinese version of TikTok, as an example, the 2025 Douyin E-commerce Development Report shows that the platform attracted more than 5.36 million new merchants and 5.11 million new creators compared with the previous year. Among them, over 80,000 newly registered merchants generated transaction volume exceeding one million Chinese yuan through livestreaming ^[2]. In addition to Douyin, major e-commerce and social media platforms such as Taobao, Jingdong, and Xiaohongshu have also made substantial investments in livestreaming. In response, many Chinese enterprises redirect their strategic focus from

the domestic market toward overseas markets. In contrast to the saturated Chinese market, the relatively slow development of the overseas livestreaming industry indicates that it still possesses considerable growth potential. Jing Daily reports that TikTok's global monthly active users surpassed 2 billion for the first time in the first quarter of 2026 ^[3]. Its large user base and interactive communication environment create new opportunities for Chinese brands to engage global consumers.

Among Chinese brands involved in cross-border livestreaming, POP MART represents a particularly distinctive case. As one of China's most influential designer toy brands, POP MART has successfully expanded its business from the domestic market to overseas markets. Xu Ying argues that Chinese designer toys, represented by POP MART, attract global consumers through unique visual design and cultural creativity ^[4]. Ran Longnan indicates

that POP MART's revenue in the Americas reached 6.81 billion yuan, representing a year-on-year increase of 748.4% [5]. These studies suggest that POP MART is not merely a toy brand but also a representative case of Chinese cultural and commercial globalization.

In line with Marketing 3.0, contemporary livestreaming is not limited to product-centered promotion but increasingly emphasizes the delivery of emotional and experiential value [6]. This tendency is particularly evident in POP MART's livestreaming practices, where suspense, surprise, and audience interaction become central to the selling process. In such livestreams, products are not merely described. Instead, they are displayed and promoted through the coordination of multiple communicative modes. Therefore, POP MART U.S. SHOP livestreams on TikTok provide a valuable context for examining how multimodal resources are organized in e-commerce communication.

Existing studies have examined cross-border livestreaming from multiple perspectives, including platform development, marketing strategies, and linguistic features. Zhang Xiying, Liu Mengwei, and Liu Min, for example, argue that Southeast Asia is becoming one of the fastest-growing regions for cross-border e-commerce due to high social media penetration, the popularity of mobile internet, and the growth of online consumption [7]. Their comparative study highlights differences between TikTok's "content-driven" traffic model and Shopee's "transaction-oriented" platform system, emphasizing that localization and entertainment are crucial to consumer engagement. Zhong Ning and Jin Fang examine the function of livestreaming in promoting rural revitalization through improving the visibility of Xinjiang agricultural products [8]. Their study emphasizes brand storytelling, regional cultural symbols, and influencer cooperation as effective methods for building consumer trust and enhancing purchase intention. Similarly, Cong Chao and Huang Junxiao investigate tropical-fruit marketing in Southeast Asia and put forward that differences in cultural cognition significantly affect audience reception, requiring differentiated script design and localized communication strategies [9]. From the perspective of linguistic features, Huang et al. examine the livestreaming discourse of Chinese live streamer Austin Li from the perspective of multimodal discourse analysis [10]. Ji Xinrui and Li Wanli analyze cross-border live streams from the auditory modality and the visual modality [11]. Ni Zhaoshu investigates pragmatic features of cross-border livestream scripts. These studies identify persuasive speech acts, emotional appeals, interactive questioning, and trust-building expressions as key components of successful communication [12].

However, several gaps remain. Current research has predominantly focused on Chinese-language livestreaming

contexts, while English-language livestreams conducted by representative Chinese brands in overseas markets have received insufficient scholarly attention. In addition, previous studies have tended to focus on platform mechanisms, marketing strategies, or linguistic features, while paying less attention to how actions are organized in real-time livestreaming interaction. In particular, how streamers coordinate speech, gaze, gesture, facial expression, posture, music, and layout to construct higher-level actions remains underexplored.

1.2 Multimodal interaction analysis and its relevance to livestreaming research

Multimodal interaction analysis, developed by Sigrid Norris, provides a theoretical framework for examining how multiple communicative modes are coordinated in human actions and interactions [13]. The framework draws theoretically on interactional linguistics, discourse analysis, and mediated discourse analysis [14–16], and has been widely recognized as a significant strand of multimodal analysis [17].

Norris classifies communicative modes into language, gesture, gaze, layout, music, and print [13], which can be further categorized into embodied modes and disembodied modes. Embodied modes refer to resources closely connected with the human body, such as language, gesture, and gaze. By contrast, disembodied modes encompass non-bodily resources, including layout, music, and print.

However, MIA does not treat communicative modes as objects of analysis. Instead, it takes actions as the basic analytical unit and examines how communicative modes are coordinated in the accomplishment of actions. In this framework, actions are understood as mediated actions, since they are carried out by social actors through various mediational means [18]. According to Norris's theory, mediated actions can be further explained in terms of lower-level actions, higher-level actions, and frozen actions. A lower-level action refers to "[t]he smallest interactional meaning unit" and is realized through a single communicative mode [13, p. 11]. A higher-level action, by contrast, "is made up of a multiplicity of chained lower-level actions" [13, p. 11]. A frozen action, which "is entailed in material objects", can function as either a higher-level or lower-level action, depending on the specific interactional context [13, p. 11].

Following Zhang and Wang's discussion of multimodal interaction analysis [19], in actual communicative activities, social actors often perform several higher-level actions simultaneously, yet the level of attention allocated to each higher-level action varies. According to the different levels of attention involved, higher-level actions can be categorized into foregrounded, mid-grounded, and

backgrounded higher-level actions. Foregrounded higher-level actions occupy the main focus of the actor's attention, whereas mid-grounded and backgrounded actions are carried out with progressively lower degrees of attention. To indicate the degree of attention allocated to a higher-level action, modal density serves as a key analytical concept. The higher the modal density of a higher-level action, the more likely the higher-level action is to be foregrounded.

According to Norris, “[m]odal density can be achieved through either modal intensity or modal complexity (or both)” [13, p. 79]. Modal intensity is related to the importance and the degree of irreplaceability of a particular mode in realizing a higher-level action. The mode demonstrating high intensity would change the higher-level action when it is discontinued. To illustrate, if spoken language ceases during a phone conversation, the higher-level action of chatting may also terminate. As Norris explains, “[t]he more intensity or weight that a mode carries, the higher the modal density” [13, p. 79]. “Modal complexity refers to the interplay of numerous communicative modes that make the construction of a higher-level action possible” [13, p. 87]. Contrary to modal intensity, a higher-level action will not terminate when a mode discontinues if the modal density is realized through modal complexity.

As a methodological framework, the modal density foreground-background continuum is further proposed [13, p.99]. **Figure 1** illustrates the modal density foreground-background continuum, where higher modal density is associated with a higher degree of attention.

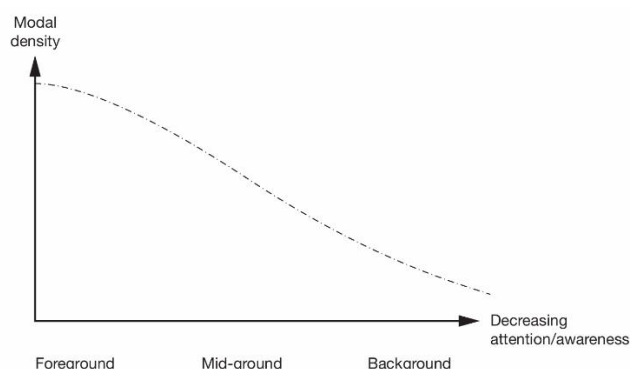


Figure 1. *The modal density foreground-background continuum.*

Existing studies have applied MIA to a range of research contexts. Zhang Yun and Li Yu explore multimodal interaction in foreign language learning [20]. Li Yuhong examines multimodal interaction patterns in multimedia-assisted college English listening classes [21]. Su Fang, Cao Min, and Li Yihui argue that MIA provides a new perspective for foreign language textbook use research [22]. Beyond educational domains, Zhu Tingting, Zhao Zhuoying, and Fan Leilei analyze the Chinese Ministry of

Foreign Affairs press conferences [23]. Guo Min and Ren Wenjuan investigate the Chinese-English publicity documentaries [24].

However, few studies explore its application in business contexts. As a distinctive interactional process, livestreaming consists of several higher-level actions: introducing the product, displaying the product, responding to audiences' comments, creating an atmosphere of urgency, and directing audiences to purchase, etc. Each of these higher-level actions is realized through chains of lower-level actions. And the degree of attention allocated to audiences also varies in different stages. These features make multimodal interaction analysis particularly suitable for live streams.

Taken together, the above discussion indicates the need to examine cross-border e-commerce livestreaming. Accordingly, this study takes POP MART U.S. SHOP livestreams on TikTok as its case and adopts multimodal interaction analysis as its analytical framework. It addresses the following research questions:

RQ1: What communicative modes are employed in different stages of POP MART's livestreams on TikTok?

RQ2: How are higher-level actions constructed through the coordination of lower-level actions across different stages of the livestreaming process?

RQ3: How does modal density contribute to the foregrounding, mid-grounding, and backgrounding of higher-level actions in cross-border livestreaming?

By answering these questions, this study extends the application of multimodal interaction analysis to e-commerce livestreaming and offers practical insights for Chinese brands seeking to optimize their communication strategies in overseas markets.

2. Materials and Methods

2.1 Data collection and sampling

This study adopts a qualitative research method to investigate multimodal interaction in POP MART U.S. SHOP livestreams on TikTok. Guided by Sigrid Norris's multimodal interaction analysis, it focuses on the higher-level actions across different livestreaming stages. The data for this study are drawn from POP MART U.S. SHOP livestreaming sessions on TikTok. The full corpus consists of three hours of livestream recordings, which are selected according to the relevance to e-commerce communication, typicality of livestreaming procedures, and richness of multimodal interaction. According to the structure of the livestreaming process, the corpus is divided into five major segments: opening, product introduction, product promotion, blind box opening, and audience engagement.

Given the large volume of livestreaming data and the micro-analytical nature of multimodal interaction analysis, this study adopts a purposive sampling strategy. From the three-hour corpus, a three-minute analytical sample is sampled for detailed examination. The selected sample covers the five aforementioned segments and contains rich multimodal resources, clear interactional actions, and observable variations in modal density and interactional attention. This sampling design allows for a focused micro-analysis of representative interactional moments in which modal density and attention allocation are particularly salient.

2.2 Data annotation

This study conducts ELAN-based annotation to identify the communicative modes in livestreaming. Drawing on MIA, communicative modes are categorized into language, posture, gaze, layout, music, and print. To improve annotation efficiency and facilitate the identification of prosodic features, we use VLC media player to extract the audio tracks from the selected livestream recordings. The video recordings and extracted audio tracks are then imported into ELAN for time-aligned annotation.

In this study, language is divided into spoken language and prosody. The former refers to the streamer's verbal linguistic information, while the latter concerns phonological features such as intonation. In addition,

posture is further specified into hand movement, head movement, facial expression, and body posture. To maintain clarity and consistency, the term eye gaze is used instead of the broader term gaze.

To improve the consistency and readability of annotation, the codes are then designed according to a “mode + subcategory” principle. For instance, PIF refers to “Prosody Intonation Fall”, HNI to “Hand Movement Iconic”. **Table 1** presents the coding scheme for multimodal annotation adopted in ELAN. Spoken language is marked as SLV, while the streamer's verbal content is transcribed in full to facilitate subsequent analysis. Since printed texts are not salient in the selected livestream segments, print is not treated as an independent annotation tier in this study.

After annotation, the distribution, duration, and co-occurrence of communicative modes are examined in relation to lower-level actions, higher-level actions, and the modal density foreground-background continuum. This procedure makes it possible to identify how lower-level actions are coordinated to construct higher-level actions across different livestreaming stages. The foregrounded, mid-grounded, and backgrounded position of each higher-level action is then determined according to the interactional weight of the modes involved, the intensity carried by particular modes, and the complexity of multimodal coordination.

Table 1. ELAN Annotation Coding Scheme ¹.

Code	Meaning	Code	Meaning
1. Spoken language /SL-/ [SLV]	Spoken language with verbal content		
2. Prosody Intonation /P-/ [PIF] [PIRF]	Intonation fall Intonation rise & fall	[PIR] [PIS]	Intonation rise Intonation stress
3. Hand movement /HN-/ [HNI] [HNDe]	Iconic Deictic	[HNM] [HNB]	Metaphoric Beats
4. Head movement /HA-/ [HAU] [HAN]	Up Nod	[HAD] [HAMA]	Down Moves around
5. Facial expression /FE-/ [FES]	Smile	[FESe]	Serious
6. Eye gaze /EG-/ [EGA]	Eye gazes at the audience	[EGAr]	Eye gazes around
7. Posture /PO-/ [POO]	Posture open	[POC]	Posture closed
8. Layout /L-/ [LYS] [LYPn]	Streamer in front Phone in front	[LYP]	Product in front
9. Music /M-/ [MUL]	Light music	[MUS]	Special music

¹ This coding scheme is adapted from previous ELAN-based multimodal annotation studies and modified according to the characteristics of POP MART's cross-border livestreaming data [25–27].

3. Results

The ELAN annotation results are presented in this section. **Table 2** illustrates the detailed temporal durations of the recorded livestream data, while **Table 3** presents a summary of the annotated data for communicative modes. According to the annotated results, spoken language, prosody, hand movement, head movement, facial expression, eye gaze, posture, music, and layout are involved in the construction

of livestreaming interaction. Spoken language occupies a prominent position, with a total duration of 148.84 seconds, accounting for 86.75% of the annotated data. Music and layout are also highly continuous, accounting for 92.69% and 100% of the total duration respectively. By contrast, prosody, head movement, facial expression, gaze, and posture appear for shorter durations, yet they often become salient at specific interactional moments.

Table 2. Live stream temporal segmentation.

Sessions of live stream	Durations
Opening segment	3.482s
Product introduction	69.1s
Product promotion	58.445s
Blind box opening	35.032s
Audience engagement	5.509s

Table 3. Summary of ELAN annotation data for communicative modes.

Tier	No. of Annotations	Average	Tot. Duration/s	Duration pct./%
Spoken language	52	24.76	148.84	86.75
Prosody	60	28.57	25.58	14.91
Hand movement	26	12.38	93.16	54.3
Head movement	16	7.62	21.66	12.62
Facial expression	13	6.19	45.14	26.31
Eye gaze	12	5.71	47.61	27.75
Posture	13	6.19	46.6	27.61
Layout	8	3.81	159.02	92.69
Music	10	4.76	171.6	100

3.1 Multimodal interactions in the opening segment of the live stream

The opening segment is the shortest segment in the selected data, lasting 3.482 seconds. Despite its brevity, it plays an important role in initiating the livestream interaction and preparing the audience for the following product-related activities. Three higher-level actions are identified in this segment: addressing the audience, directing the audience's attention, and maintaining a commercial atmosphere. Among them, addressing the audience is foregrounded, while directing the audience's attention is positioned in the mid-ground and maintaining a commercial atmosphere remains backgrounded.

As shown in **Table 4**, the streamer produces five instances of spoken language, lasting 1.23 seconds and accounting for 35% of the segment. Prosodic features are also present, including intonation fall, intonation rise, and prosodic stress. Hand movement consists of one deictic gesture and one beat gesture, accounting for 52.2% and 47.8% of the segment respectively. Although the verbal content is relatively brief, several other modes continue throughout the whole segment. Eye gaze at the audience, serious facial expression, streamer-in-front layout, and light music each occupy 100% of the segment, providing a stable visual and auditory basis for the opening interaction.

Table 4. ELAN annotation results for the opening segment.

Tier	Annotation content	No. of annotations	Duration/s	Duration pct./%
	Opening	Opening	Opening	Opening
Spoken language	SLV	5	1.23	35
Prosody	PIF	2	0.84	24
	PIR	1	0.12	3.6
	PIS	2	0.26	7.6

Tier	Annotation content	No. of annotations	Duration/s	Duration pct./%
	Opening	Opening	Opening	Opening
Hand movement	HNDe	1	1.82	52.2
	HNB	1	1.66	47.8
Head movement	HAU	1	0.12	3.5
Facial expression	FESe	1	3.48	100
Eye gaze	EGA	1	3.48	100
	POO	1	1.95	56.1
Posture	POC	1	1.05	30.3
Layout	LYS	1	3.48	100
Music	MUL	1	3.48	100

Table 5 further summarizes the higher-level actions and their positions on the modal density foreground-background continuum. Addressing the audience is constructed through the coordination of spoken language, prosody, beat gesture, eye gaze, facial expression, posture, layout, and music. Directing the audience's attention is mainly supported by deictic hand movement, spoken language, prosodic stress, eye gaze, and layout. By contrast, maintaining a commercial atmosphere is supported

primarily by streamer-in-front layout and light music. These results indicate that the opening segment does not rely on verbal greeting alone. Instead, it establishes initial audience contact through a compact coordination of speech, gaze, facial expression, gesture, posture, layout, and music. The ELAN annotation interface in **Figure 2** presents the time-aligned organization of these modes in the opening segment.

Table 5. Modal density foreground-background continuum in the opening segment.

Higher-level action	Communicative modes	Modal density	Continuum position
Addressing the audience	SLV, PIF, PIR, PIS, HNB, EGA, FESe, POO, LYS, MUL	High	Foreground
Directing the audience's attention	HNDe, SLV, PIS, EGA, LYS	Medium-high	Mid-ground
Maintaining a commercial atmosphere	LYS, MUL	Low-medium	Background

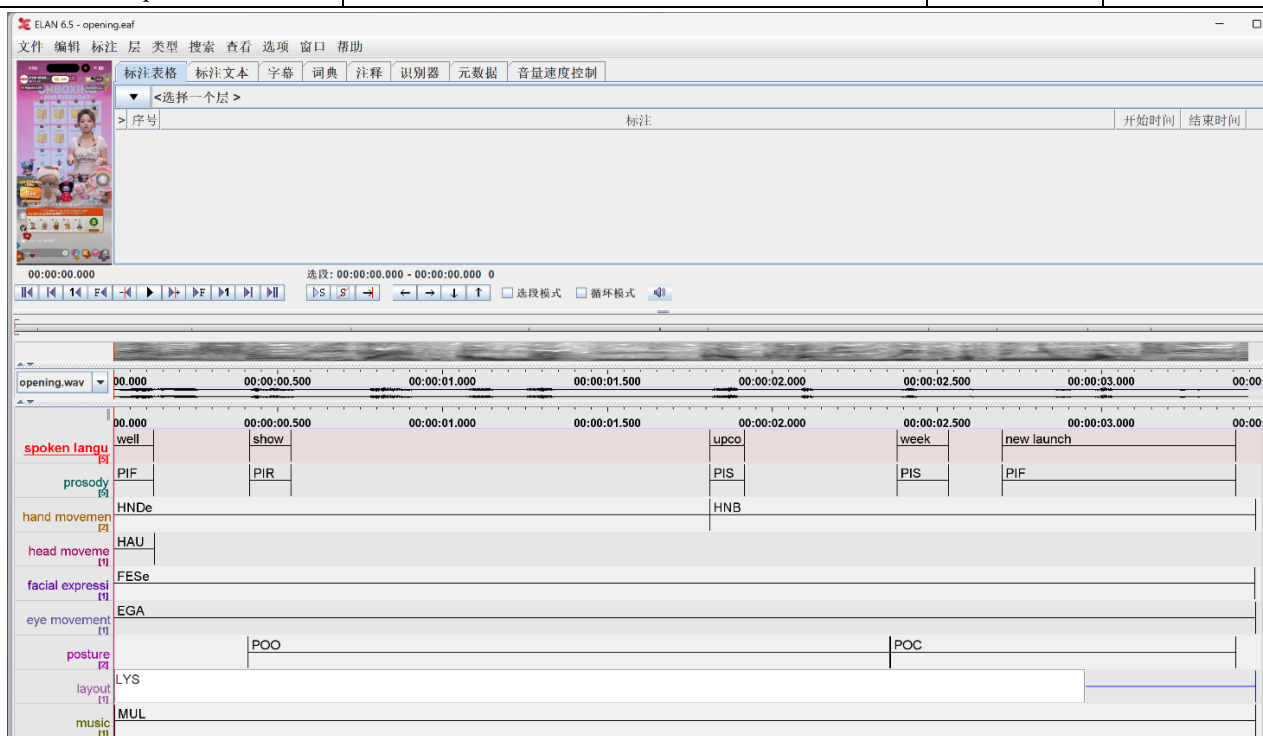


Figure 2. ELAN annotation interface of the opening segment.

3.2 Multimodal interactions in the product introduction segment of the live stream

The product introduction segment is the longest part of the selected data, lasting 69.1 seconds. Its central higher-level action is introducing the product, while displaying the product, directing the audience's attention to the product, and maintaining the livestream atmosphere function as supporting actions. As shown in **Table 6**, spoken language is almost continuous in this segment, lasting 69.08 seconds

and accounting for 99.4% of the segment. The product-in-front layout also continues throughout the whole segment, providing a stable visual focus for the introduction. In addition, deictic hand movement lasts 44.38 seconds, accounting for 64.2%, which shows that the streamer frequently uses hand movement to orient the audience toward the product being introduced. Prosodic features, especially intonation rise and prosodic stress, occur intermittently and help mark selected information in the verbal explanation.

Table 6. ELAN annotation results for the product introduction segment.

Tier	Annotation content	No. of annotations	Duration/s	Duration pct./%
	Product introduction	Product introduction	Product introduction	Product introduction
Spoken language	SLV	3	69.08	99.4
Prosody	PIF	1	0.09	0.1
	PIR	16	2.72	3.9
	PIS	7	1.54	2.2
Hand movement	HNDe	2	44.38	64.2
Layout	LYP	1	69.1	100
Music	MUL	1	69.1	100

Table 7 further presents the positions of the higher-level actions on the modal density foreground-background continuum. Introducing the product is foregrounded because it is sustained by the continuous coordination of spoken language, product display, deictic hand movement, and selected prosodic features. Displaying the product and directing the audience's attention to the product are

positioned in the mid-ground, as they support and strengthen the product introduction rather than replace it as the central action. Maintaining the commercial livestream atmosphere remains backgrounded and is mainly supported by light music. The ELAN annotation interface in **Figure 3** visually presents the time-aligned organization of these modes in the product introduction segment.

Table 7. Modal density foreground-background continuum in the product introduction segment.

Higher-level action	Communicative modes	Modal density	Continuum position
Introducing the product	SLV, LYP, HNDe, PIR, PIS, MUL	High	Foreground
Displaying the product	HNDe, LYP	Medium-high	Mid-ground
Directing the audience's attention	SLV, PIR, PIS	Medium	Mid-ground
Maintaining the commercial live stream atmosphere	MUL	Low-medium	Background

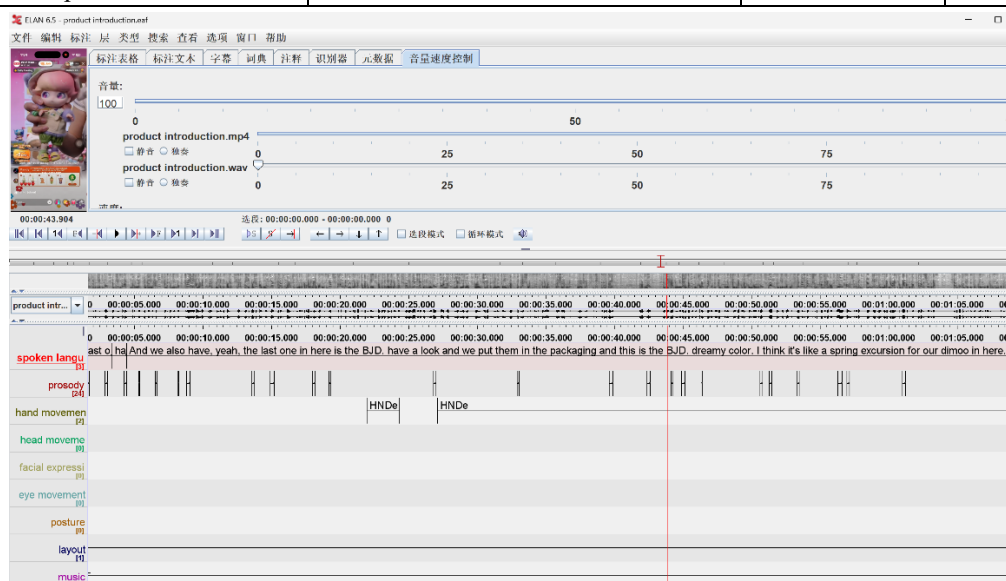


Figure 3. ELAN annotation interface of the product introduction segment.

3.3 Multimodal interactions in the product promotion segment of the live stream

The product promotion segment differs from the product introduction segment in both interactional focus and modal organization. Rather than mainly providing product information, this segment is oriented toward persuasion and purchase guidance. Five higher-level actions are identified: promoting the product and persuading audiences to purchase, constructing product value, directing audiences' attention to purchasing information, maintaining interpersonal engagement, and maintaining the livestream atmosphere.

The annotation results in **Table 8** show that spoken language remains the major verbal resource, lasting 50.9 seconds and accounting for 87.1% of the segment. However, the promotional segment involves a richer combination of modes than the product introduction segment. Prosodic stress appears 11 times, more frequently than in the other segments, suggesting that the streamer repeatedly emphasizes key promotional points. Hand movement also becomes more varied. Metaphoric gestures last 15.65 seconds, deictic gestures last 9.66 seconds, and beat gestures last 7.37 seconds. Eye gaze at the audience, serious facial expression, and open posture each occupy 50.6% of the segment, indicating that the streamer maintains a strong audience-oriented stance while promoting the product. In terms of layout, both streamer-in-front layout and phone-in-front layout are used, with the latter making purchasing-related information visually available to the audience.

Table 8. ELAN annotation results for the product promotion segment.

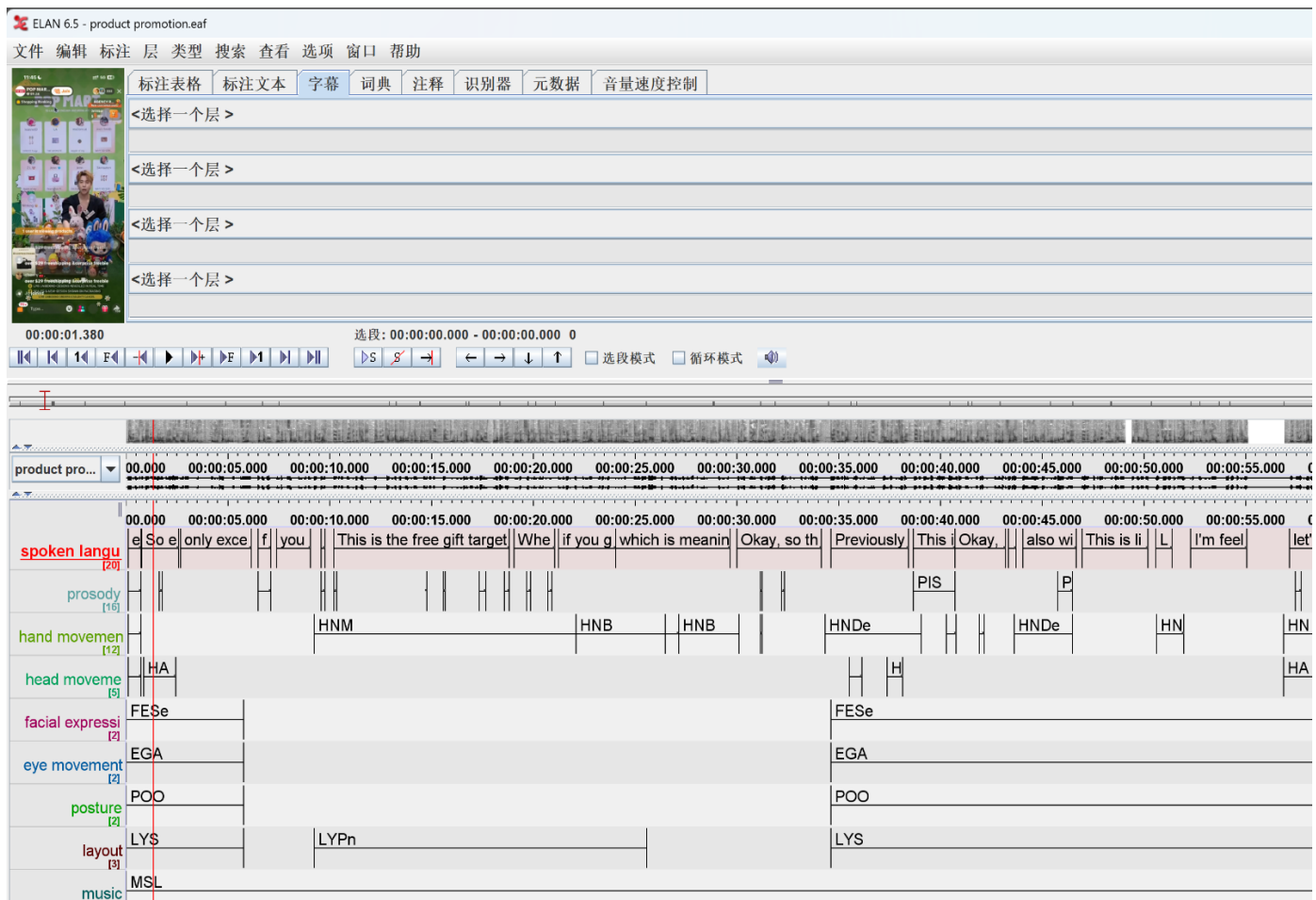
Tier	Annotation content	No. of annotations	Duration/s	Duration pct./%
	Product promotion	Product promotion	Product promotion	Product promotion
Spoken language	SLV	20	50.9	87.1
Prosody	PIR	4	1.42	2.4
	PIRF	1	0.62	1.1
	PIS	11	3.96	6.8
Hand movement	HNM	5	15.65	26.8
	HNDe	5	9.66	16.5
	HNB	2	7.37	12.6
Head movement	HAU	1	1.57	2.7
	HAD	1	0.62	1.1
	HAMA	2	2.26	3.9
	HAN	1	0.62	1.1
Facial expression	FESe	2	29.59	50.6
Eye gaze	EGA	2	29.59	50.6
Posture	POO	2	29.59	50.6
Layout	LYS	2	29.58	50.6
	LYPn	1	16.35	28
Music	MUL	1	58.44	100

As summarized in **Table 9**, promoting the product and persuading audiences to purchase are positioned in the foreground. This foregrounded action is supported not only by spoken language but also by prosodic stress, metaphoric and deictic gestures, gaze, facial expression, posture, layout, and music. Constructing product value and directing audiences' attention to purchasing information are placed

between the foreground and mid-ground because they directly support the persuasive purpose of the segment. Maintaining interpersonal engagement is positioned in the mid-ground, while maintaining the livestream atmosphere remains in the background. **Figure 4** shows the ELAN annotation interface of this segment and illustrates how these modes are distributed across the promotional process.

Table 9. Modal density foreground-background continuum in the product promotion segment.

Higher-level action	Communicative modes	Modal density	Continuum position
Promoting the product and persuading audiences to purchase	SLV, PIS, HNMM, HNDe, HNB, EGA, FESe, POO, LYS, LYPn, MUL	High	Foreground
Constructing product value	SLV, HNMM, PIS, FESe, POO	Medium-high	Foreground & mid-ground
Directing audiences' attention to purchasing information	HNDe, LYPn, SLV, EGA, PIS	Medium-high	Mid-ground & foreground
Maintaining interpersonal engagement	EGA, FESe, POO, SLV	Medium	Mid-ground
Maintaining the live stream atmosphere	MUL	Low-medium	Background

**Figure 4.** ELAN annotation interface of the product promotion segment.

3.4 Multimodal interactions in the blind box opening segment of the live stream

The blind box opening segment presents a different multimodal pattern from the previous two segments. It lasts approximately 35 seconds and is organized around suspense, reveal, and evaluation. Five higher-level actions are involved: opening the blind box, revealing the hidden item, evaluating the revealed product, maintaining audience engagement, and maintaining the livestream atmosphere. Among these, opening the blind box and revealing the hidden item are foregrounded.

As shown in **Table 10**, spoken language appears 18 times,

lasting 22.99 seconds and accounting for 65.4% of the segment. Compared with the product introduction segment, spoken language remains important but is not overwhelmingly dominant. Prosodic features become more active in this segment: prosodic stress lasts 5.40 seconds, intonation rise and fall lasts 3.58 seconds, and intonation rise lasts 1.11 seconds. These prosodic changes help shape the rhythm of suspense and anticipation. Head movement, facial expression, eye gaze, and music also participate in the blind box opening event. In particular, special music lasts 12.6 seconds, accounting for 36% of the segment, and works together with the full-screen reveal animation to mark the reveal moment as interactionally salient.

Table 10. ELAN annotation results for the blind box opening segment.

Tier	Annotation content	No. of annotations	Duration/s	Duration pct./%
	Blind box opening	Blind box opening	Blind box opening	Blind box opening
Spoken language	SLV	18	22.99	65.4
Prosody	PIR	1	1.11	3.2
	PIRF	3	3.58	10.2
	PIS	6	5.40	15.4
Hand movement	HNH	6	8.36	23.9
	HNB	1	2.14	6.1
Head movement	HAU	6	8.57	24.5
	HAD	2	2.92	8.3
Facial expression	FES	3	1.92	5.5
	FESe	4	5.83	16.6
Eye gaze	EGA	5	10.47	29.9
Posture	POO	1	7.86	22.4
	POC	4	1.17	3.3
Layout	LYS	1	35	100
Music	MUL	3	22.47	64.1
	MUS	2	12.6	36

Table 11 shows that opening the blind box and revealing the hidden item are both positioned in the foreground. The former is constructed through the coordination of spoken language, prosody, hand movement, head movement, facial expression, gaze, posture, layout, and music. The latter is foregrounded through a more concentrated visual and auditory shift, especially the full-screen reveal animation and special music.

Evaluating the revealed product and maintaining audience engagement are positioned in the mid-ground, as they support the reveal without becoming the central action. Maintaining the livestream atmosphere remains backgrounded. The ELAN annotation interface in **Figure 5** presents the multimodal organization of this segment.

Table 11. Modal density foreground-background continuum in the blind box opening segment.

Higher-level action	Communicative modes	Modal density	Continuum position
Opening the blind box	SLV, PIS, PIRF, HNM, HAU, HAD, FESe, EGA, POO, LYS, MUL, MUS	High	Foreground
Revealing the hidden item	Full-screen reveal animation; MUS	High	Foreground
Evaluating the revealed product	SLV, PIS, HNM, FES, FESe, HAU, HAD	Medium	Mid-ground
Maintaining audience engagement	EGA, SLV, PIR, PIRF, FES, FESe	Medium	Mid-ground
Maintaining the live stream atmosphere	LYS, MUL	Low-medium	Background

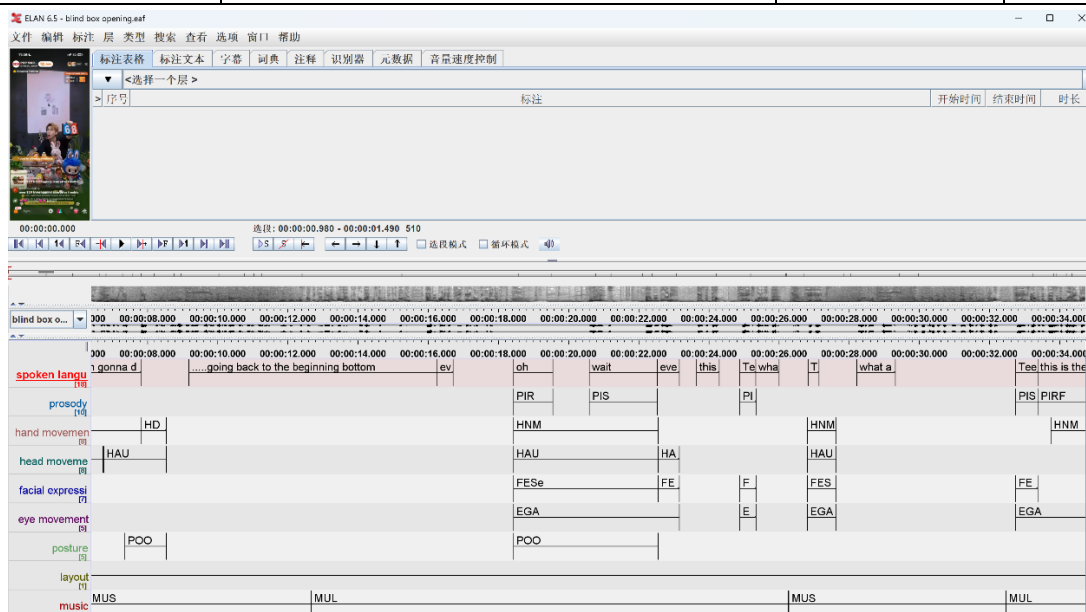


Figure 5. ELAN annotation interface of the blind box opening segment.

3.5 Multimodal interactions in the audience engagement segment of the live stream

The audience engagement segment consists of two selected clips, and their annotation results are combined for analysis. Although this segment lasts only 5.51 seconds, it contains dense multimodal coordination. Three higher-level actions are identified: engaging the audience, responding to the audience, and maintaining the livestream atmosphere.

Table 12 shows that spoken language lasts 4.64 seconds, accounting for 84.3% of the segment. Prosodic features are

also active within this short period, especially intonation rise and fall and prosodic stress. Head movement around lasts 4.98 seconds, accounting for 90.3%, which suggests that the streamer shifts attention across the screen, the camera, and the interactive environment. Eye gaze at the audience lasts 3.49 seconds, while gaze around appears briefly. Facial expression shifts between smile and seriousness, and posture alternates between open and closed positions. Streamer-in-front layout and light music continue throughout the segment, providing a stable visual and auditory background.

Table 12. ELAN annotation results for the audience engagement segment.

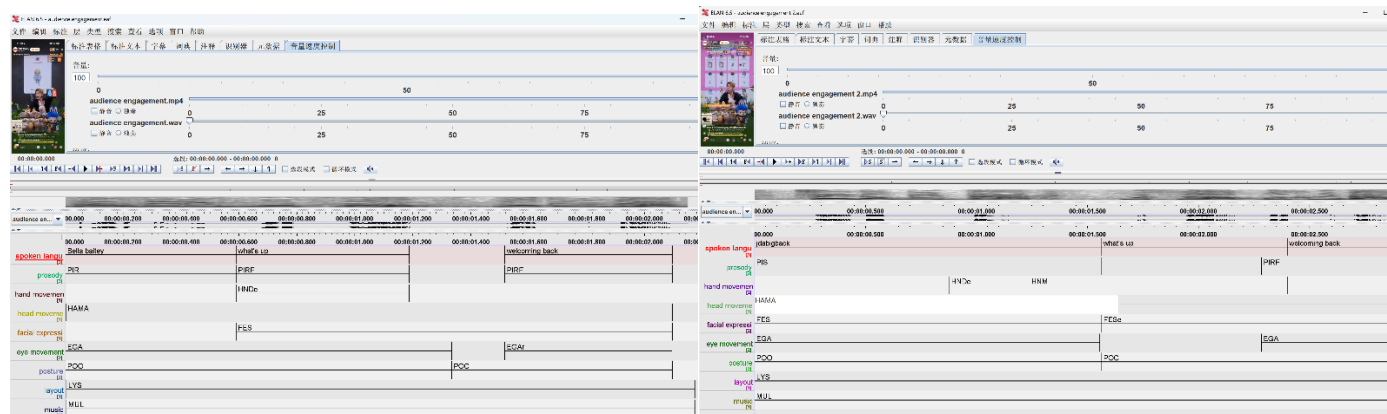
Tier	Annotation content	No. of annotations	Duration/s	Duration pct./%
	Audience engagement	Audience engagement	Audience engagement	Audience engagement
Spoken language	SLV	6	4.64	84.3
Prosody	PIR	1	0.59	10.8
	PIRF	3	1.77	32.1
	PIS	1	1.56	28.4
Hand movement	HNМ	1	1.16	21.1
	HNDe	2	0.96	17.4
Head movement	HAMA	2	4.98	90.3
Facial expression	FES	2	3.01	54.6
	FESe	1	1.31	23.8
Eye gaze	EGA	3	3.49	63.3
	EGAr	1	0.58	10.5
Posture	POO	2	2.9	52.7
	POC	2	2.08	37.7
Layout	LYS	2	5.51	100
Music	MUL	2	5.51	100

As presented in **Table 13**, engaging the audience is foregrounded through the coordination of spoken language, prosody, gaze, facial expression, posture, head movement, and hand movement. Responding to audiences is positioned in the mid-ground because it supports audience engagement by creating a sense of real-time responsiveness. Maintaining the livestream atmosphere remains

backgrounded and is mainly supported by the continuous presence of layout and light music. Since the audience engagement segment consists of two selected clips, **Figure 6** presents both ELAN annotation interfaces: Figure 6 (A) shows the first selected clip, while Figure 6 (B) shows the second selected clip.

Table 13. Modal density foreground-background continuum in the audience engagement segment.

Higher-level action	Communicative modes	Modal density	Continuum position
Engaging the audience	SLV, PIRF, PIS, EGA, FES, POO, HAMA, HNМ, HNDe	High	Foreground
Responding to audiences	HAMA, EGA, EGAr, SLV, PIRF, PIS	Medium-high	Mid-ground
Maintaining the live stream atmosphere	LYS, MUL	Low-medium	Background



(A)

(B)

Figure 6. ELAN annotation interfaces of the audience engagement segment. (A) First selected clip; (B) second selected clip.

4. Discussion

The results reveal that POP MART's TikTok livestreaming is organized as a staged multimodal activity rather than as a continuous process of product explanation. Each segment has its own communicative orientation, and the streamer selects and combines modes according to the action that needs to be accomplished at that moment. In this sense, livestreaming interaction is not a simple accumulation of isolated communicative modes. It is better understood as the coordination of multiple lower-level actions that together construct foregrounded, mid-grounded, and backgrounded higher-level actions.

The first important finding concerns the selective use of communicative modes across livestreaming stages. Spoken language appears across all five segments, but its function changes from one stage to another. In the product introduction segment, it serves as the main vehicle for information delivery. The streamer relies on continuous verbal explanation to present product features, while the product-in-front layout and deictic hand movement provide visual and embodied support. In the product promotion segment, spoken language becomes more persuasive. It is no longer used only to describe the product, but also to highlight value, create purchase motivation, and direct audiences toward buying-related information. In the blind box opening segment, spoken language becomes less dominant, as suspense and emotional involvement are increasingly carried by prosody, facial expression, music, and visual effects. This shift shows that the same mode may serve different interactional purposes depending on the stage of the livestream. In addition, embodied modes also play different roles across the five segments. In the opening segment, gaze and facial expression help establish initial contact with the audience. In the product introduction segment, deictic hand movement guides attention to the product and supports verbal explanation. In the product promotion segment, metaphoric and deictic gestures become more active because the streamer needs not only to

present the product but also to construct value and encourage purchase. In the audience engagement segment, head movement, gaze, smile, and posture help create a sense of real-time responsiveness. These changes indicate that embodied modes are not decorative additions to speech. Rather, they participate directly in the organization of livestreaming actions.

The second finding involves the construction of higher-level actions. In each segment, several higher-level actions are carried out simultaneously, but they do not receive equal interactional attention. In the opening segment, addressing the audience is foregrounded, while directing attention and maintaining the commercial atmosphere remain supportive. In the product introduction segment, introducing the product is foregrounded, with displaying the product and directing the audience's attention positioned in the mid-ground. In the product promotion segment, promoting the product and persuading audiences to purchase form the central action, while constructing product value and directing attention to purchasing information work as closely related supporting actions. In the blind box opening segment, opening the blind box and revealing the hidden item become foregrounded because the interaction is organized around suspense and surprise. In the audience engagement segment, engaging the audience is foregrounded, while responding to audience comments helps sustain the sense of interaction. This layered organization shows that livestreaming interaction is not linear. The streamer does not simply move from one isolated action to another. Instead, several actions are constructed at the same time, but with different degrees of attention. For example, during product promotion, the streamer may verbally introduce product value, use gesture to emphasize key points, look toward the audience, maintain an open posture, and display the phone or product on screen. These lower-level actions jointly construct the foregrounded promotional action, while also supporting interpersonal engagement and purchase guidance. The

interaction therefore depends on the coordination of actions at different levels.

The third finding is related to modal density. The analysis shows that modal density can be produced in different ways. In the product introduction segment, modal density is achieved mainly through modal intensity. Spoken language lasts almost throughout the segment and carries strong interactional weight. If the verbal explanation is removed, the action of introducing the product would be substantially weakened. The product-in-front layout also carries strong intensity because it keeps the product visually available and the live streamers verbal explanation when displaying the product. In contrast, the opening and audience engagement segments show that modal density may also be achieved through modal complexity. These segments are short, but they involve the simultaneous coordination of speech, prosody, gaze, facial expression, gesture, posture, layout, and music. No single mode alone determines the foregrounded action. Instead, the salience of audience-addressing and audience engagement emerges from the combined operation of several modes within a limited period of time. This is why a short segment can still be interactionally dense. The product promotion segment combines both modal intensity and modal complexity. Spoken language provides sustained persuasive content, while prosodic stress highlights selected points. At the same time, metaphoric gesture, deictic gesture, gaze, facial expression, posture, phone-in-front layout, and music together create a purchase-oriented environment. The foregrounding of promotion is therefore not achieved by speech alone, but by the way speech is intensified, embodied, visually supported, and interactionally directed toward the audience. The blind box opening segment further demonstrates that modal density is closely related to timing. The reveal action does not occupy the longest duration, but it becomes highly foregrounded because full-screen animation and special music occur at the key moment of the livestream. These modes temporarily interrupt the ordinary visual and auditory background and redirect attention to the hidden item. This suggests that short-duration modes can carry strong interactional weight when they appear at the right moment. On the contrary, modes such as light music and stable layout may last throughout a segment but remain backgrounded if they mainly provide atmosphere rather than organize the central action. In addition, mode duration alone does not determine modal density. What matters more is whether a mode is interactionally weighted, whether it is replaceable, how it coordinates with other modes, and when it appears in the unfolding interaction. This finding is important for understanding e-commerce livestreaming because livestreams often contain many simultaneous modes. A mode may be constantly present but interactionally

peripheral, while another may appear only briefly but become central to the foregrounding of a higher-level action.

Theoretically, this study extends the application of multimodal interaction analysis to the field of e-commerce livestreaming. Previous applications of this framework have mainly focused on educational communication, classroom interaction, public discourse, and media discourse. By applying the concepts of lower-level actions, higher-level actions, modal density, modal intensity, modal complexity, and the foreground-background continuum to POP MART's TikTok livestreams, this study demonstrates the explanatory value of multimodal interaction analysis in e-commercial context. It also shows that livestreaming should not be understood merely as product presentation or marketing promotion, but as a complex multimodal interactional process.

These findings also have practical implications for cross-border e-commerce livestreaming. For streamers, effective communication does not depend on adding more modes indiscriminately. Speech, gesture, gaze, facial expression, music, product display, and screen-based resources need to be coordinated with the action being performed. Product introduction requires clear verbal explanation and stable visual display. Product promotion requires persuasive speech, prosodic emphasis, gesture, and purchase-oriented visual guidance. Blind box opening requires suspense-building and a carefully timed reveal. Audience engagement requires gaze, facial expression, posture, and responsive verbal interaction. If modes are used without a clear interactional purpose, they may create noise rather than strengthen communication. For Chinese brands operating in overseas markets, this suggests that livestreaming strategies should be action-oriented. Streamers need to identify the main action at each moment and decide which modes should be foregrounded, which should remain supportive, and which should simply maintain the background atmosphere. Such coordination can make livestreaming communication clearer, more emotionally engaging, and more persuasive for international audiences.

However, several limitations should be acknowledged. The data are drawn from a single brand whose livestreaming practices are closely associated with designer toy culture and blind box consumption. Therefore, the findings may not fully represent livestreaming practices in other product categories or platform contexts. In addition, this study adopts a qualitative approach and focuses on selected high-density segments, which allows for detailed micro-analysis but may not capture the full variability of multimodal interaction across entire livestreaming sessions. Although ELAN-based annotation provides a systematic analytical

procedure, the interpretation of multimodal features inevitably involves a certain degree of subjectivity. Moreover, this study mainly focuses on the live streamer's multimodal behavior, while audience responses are discussed only through observable interactional cues.

Future research may expand the data by including more brands, platforms, product categories, and livestreaming contexts. Comparative studies across different types of products, such as cosmetics, fashion, electronics, or agricultural products, may further reveal whether the analysis of higher-level actions identified in this study are generalizable to other forms of e-commerce livestreaming. Future studies may also combine qualitative multimodal analysis with quantitative methods, such as comment analysis, audience engagement metrics, or purchasing data, in order to examine more directly the relationship between multimodal interaction and consumer response. Given the cross-border context of POP MART's TikTok livestreams, future research may also adopt a more audience-oriented and cross-cultural perspective to explore how multimodal strategies are adapted to different linguistic, cultural, and market environments.

Author Contributions

Conceptualization, Xiangyang Zhang; methodology, Xiangyang Zhang; data collection, Xiangyang Zhang; data annotation and analysis, Xiangyang Zhang; writing—original draft preparation, Xiangyang Zhang; writing—review and editing, Xiangyang Zhang; supervision, Xiangyang Zhang. All authors have read and agreed to the published version of the manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

This work received no external funding.

Institutional Review Board Statement

Ethical review and approval were waived for this study because it analyzes publicly available livestreaming materials and does not involve direct interaction with human participants, intervention, or the collection of private personal information.

Informed Consent Statement

Not applicable. The study analyzes publicly available livestreaming materials for academic purposes and does not involve direct recruitment, intervention, interviews, surveys, or the collection of private personal information from human participants.

Data Availability Statement

The data used in this study are collected by the author from publicly accessible POP MART U.S. SHOP livestreams on

TikTok. Due to platform-related restrictions, copyright considerations, and the presence of identifiable streamer images and voices, the raw video recordings and ELAN annotation files are not publicly available. Relevant analytical descriptions and summarized annotation results are provided in the article. Anonymized or summarized annotation materials may be made available from the corresponding author upon reasonable request, where permitted by platform policies and copyright restrictions.

Acknowledgements

The authors would like to thank all those who provided support during the preparation of this manuscript.

References

1. Buchholz, K. (2021). This chart shows e-commerce is growing fastest in China. World Economic Forum. Available from: <https://www.weforum.org/stories/2021/01/ecommerce-sales-china-global-growth/> (cited 11 June 2026).
2. Xinhua News Agency. (2026). 抖音电商年度报告: 超8万新商家直播成交额破百万 [Douyin E-commerce annual report: More than 80,000 new merchants achieved livestreaming transaction volume exceeding one million yuan]. Xinhuanet. [in Chinese]. Available from: <https://www.news.cn/tech/20260115/4edff26a18614c088844c150025f8c89/c.html> (cited 11 June 2026).
3. Jing Daily. (2026). TikTok tops 2 billion users, shop GMV climbs 94% in 2025. Jing Daily. Available from: <https://jingdaily.com/intels/2026-04/29/tiktok-tops-2-billion-users-shop-gmv-climbs-94-in-2025> (cited 11 June 2026).
4. Xu, Y. (2025). The Labubu effect: Chinese designer toys are sparking a global craze with bold designs and cultural flair. *ChinAfrica*. 17(11), 56–57.
5. Ran, L. N. (2026). 增长20%泡泡玛特定下2026年目标 [POP MART sets its 2026 target after 20% growth]. *China Business Journal*. 31 March 2026, 004. [in Chinese]. DOI: <https://doi.org/10.38304/n.cnki.nzgsb.2026.000565>
6. Kotler, P., Kartajaya, H., & Setiawan, I. (2010). *Marketing 3.0: From Products to Customers to the Human Spirit*. John Wiley & Sons: Hoboken, NJ, USA.
7. Zhang, X. Y., Liu, M. W., & Liu, M. (2023). 跨境电商直播在东南亚市场的发展现状及对策—

—基于TikTok和Shopee的对比分析 [The development status and countermeasures of cross-border e-commerce livestreaming in the Southeast Asian market: A comparative analysis based on TikTok and Shopee]. *Times Economy and Trade*. 20(8), 94–96. [in Chinese]. DOI: <https://doi.org/10.19463/j.cnki.sjtm.2023.08.012>

8. Zhong, N., & Jin, F. (2026). 乡村振兴背景下新疆农特产品直播电商营销策略研究 [A study on livestreaming e-commerce marketing strategies for Xinjiang agricultural specialty products in the context of rural revitalization]. *Shanxi Agricultural Economy*. (4), 202–204, 230. [in Chinese]. DOI: <https://doi.org/10.16675/j.cnki.cn14-1065/f.2026.04.061>
9. Cong, C., & Huang, J. X. (2026). 基于文化认知差异的AIGC直播脚本优化研究——以东南亚热带水果跨境营销为例 [A study on the optimization of AIGC livestreaming scripts based on cultural cognitive differences: A case study of cross-border marketing of tropical fruits in Southeast Asia]. *Market Modernization*. (6), 84–86. [in Chinese]. DOI: <https://doi.org/10.14013/j.cnki.scxdh.2026.06.048>
10. 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: Stephanidis, C. (ed.). *HCI International 2020—Late Breaking Papers: Interaction, Knowledge and Social Media*. Lecture Notes in Computer Science, vol. 12427. Springer: Cham, Switzerland. pp. 305–327. DOI: https://doi.org/10.1007/978-3-030-60152-2_24
11. Ji, X. R., & Li, W. L. (2021). 多模态视角下跨境电商直播话语分析 [A discourse analysis of cross-border e-commerce livestreaming from a multimodal perspective]. *Overseas English*. (20), 92–93. [in Chinese].
12. Ni, Z. S. (2024). 跨境电商直播话术的语用策略浅析 [An analysis of pragmatic strategies in cross-border e-commerce livestreaming scripts]. *Vocational Education*. 23(4), 29–33. [in Chinese]. DOI: <https://doi.org/10.3969/j.issn.2095-4530.2024.04.008>
13. Norris, S. (2004). *Analyzing Multimodal Interaction: A Methodological Framework*. Routledge: New York, NY, USA.
14. Stivers, T., & Sidnell, J. (2005). Introduction: multimodal interaction. *Semiotica*. 156(1/4), 1–20. DOI: <https://doi.org/10.1515/semi.2005.2005.156.1>
15. [15] Harris, Z. S. (1952). Discourse analysis. *Language*. 28(1), 1–30. DOI: <https://doi.org/10.2307/409987>
16. Scollon, R. (2001). *Mediated Discourse: The Nexus of Practices*. Routledge: London, UK.
17. Jewitt, C. (2009). Different approaches to multimodality. In: Jewitt, C. (ed.). *The Routledge Handbook of Multimodal Analysis*. Routledge: London and New York. pp. 28–39.
18. Norris, S., & Jones, R. H. (eds.). (2005). *Discourse in Action: Introducing Mediated Discourse Analysis*. Routledge: London, UK.
19. Zhang, D. L., & Wang, Z. (2016). 多模态互动分析框架探索 [Exploring a framework of multimodal interaction analysis]. *China Foreign Languages*. 13(2), 54–61. [in Chinese]. DOI: <https://doi.org/10.13564/j.cnki.issn.1672-9382.2016.02.008>
20. [20] Zhang, Y., & Li, Y. (2012). 多模态互动外语学习探究——教学设计个案 [A study on multimodal interactive foreign language learning: A case of instructional design]. *Continuing Education Research*. (11), 171–172. [in Chinese]. DOI: <https://doi.org/10.3969/j.issn.1009-4156.2012.11.059>
21. Li, Y. H. (2013). 多媒体环境下大学英语听力教学中的多模态互动模式 [A multimodal interaction model in college English listening teaching in a multimedia environment]. *The Youth Writers*. (12), 68–69. [in Chinese].
22. Su, F., Cao, M., & Li, Y. H. (2025). 外语教材使用研究新视角——多模态互动分析 [A new perspective on foreign language textbook use research: Multimodal interaction analysis]. *Foreign Language Textbook Research*. (2), 114–127. [in Chinese].
23. Zhu, T. T., Zhao, Z. Y., & Fan, L. L. (2025). 中国外交部新闻发言人例行记者会多模态互动话语研究 [A multimodal interaction discourse study of regular press conferences by spokespersons of the Chinese Ministry of Foreign Affairs]. *Journalism and Communication*. (22), 45–47. [in Chinese].

24. Guo, M., & Ren, W. J. (2025). 基于多模态话语分析的中英外宣纪录片模态互动机制研究 [A study on the modal interaction mechanism of Chinese and English publicity documentaries based on multimodal discourse analysis]. *Journal of Xinyang Normal University (Philosophy and Social Sciences Edition)*. 45(5), 97–105. [in Chinese]. DOI: <https://doi.org/10.3969/j.issn.2097-5821.2025.05.015>
25. Wang, L. F., & Wen, Y. (2008). 应用语言学的多模态分析方法 [Multimodal analysis methods in applied linguistics research]. *Technology Enhanced Foreign Language Education*. (3), 8–12. [in Chinese]. DOI: <https://doi.org/10.3969/j.issn.1001-5795.2008.03.002>
26. Guan, Y. (2021). Multimodal interaction mechanism of speech videos based on ELAN. *Journal of University of Science and Technology Beijing (Social Sciences Edition)*. 37(5), 499–510. [in Chinese].
27. McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press: Chicago, IL, USA.