

What a Multimodal Information Density Shapes Audience Engagement: Evidence from Sports Event Mascot Short-Form Videos

YuKun Li

Chengdu University of Technology

Yu Shan

shangyu15@cdut.edu.cn

Chengdu University of Technology

ZhengHao Peng

Chengdu University of Technology

Ping Nie

Chengdu University of Technology

Hong Wang

Chengdu University of Technology

Article

Keywords: Multimodal information density, Sports event mascots, Short-form video engagement, Cognitive load theory, Content source type

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11059339/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

As short-form video emerges as a dominant format in global digital culture, sports event mascots have evolved from static symbolic emblems into dynamic multimodal digital content circulating on youth-oriented platforms. While prior scholarship has explored mascot design, cultural symbolism, and audience engagement, how the intensity of multimodal information embedded in short videos shapes user responses remains underexplored, particularly regarding nonlinear effects and the moderating role of content origin. Grounded in cognitive load theory and computational communication analysis, this study examines how visual, textual, and audio information density influence users' liking behavior toward sports mascot short videos. Based on 173 mascot-centered videos collected from Bilibili, we construct three modality-specific density indicators via a systematic feature engineering pipeline and test their linear, nonlinear, and source-contingent effects using hierarchical regression models. Results show all three density dimensions are positively associated with liking, with textual density exerting the strongest effect. Each dimension exhibits an inverted U-shaped relationship, with visual density showing the strongest negative nonlinear component. Video source type significantly moderates these patterns: textual density has stronger impact in Officially Generated Content, while audio and visual density dominate in User Generated Content. These findings extend mascot communication research, validate cognitive load boundary conditions in dynamic media contexts, and demonstrate the value of computational approaches for audiovisual cultural content analysis.

Introduction

The rapid growth of short-form video platforms has fundamentally transformed how consumers encounter, interpret, and interact with digital content (Dong et al. 2024). Short-form video has evolved into a dominant format in global digital media consumption. As of 2025, 87.5% of online adults worldwide watch short-form video on a weekly basis, and video constitutes the largest category of global internet traffic, with short-form video serving as the primary growth driver (We Are Social and Meltwater 2025; AppLogic Networks 2025). Within sports digital communication, sports-related IP content, including event mascot videos, has gained increasing visibility on youth-oriented video communities such as Bilibili (Bilibili Inc. 2026; Chen et al. 2021). Among these symbolic IP assets, sports event mascots have evolved beyond static visual emblems of tournaments into dynamic digital content circulated through social-media platforms (Smith et al. 2012; Kolotouchkina 2018). Unlike traditional mascot communication that mainly relies on static visual design and symbolic connotations, modern mascot short videos integrate visual footage, textual captions, subtitles, bullet comments, sound effects, and background music within a limited viewing window (Dong et al. 2024; Xiao et al. 2025). This transformation raises a critical research question for digital sports marketing: how does multimodal information density embedded in sports-event mascot short videos shape users' liking behavior?

Existing scholarship on sports-event mascots has examined anthropomorphism, symbolic representation, cultural meaning, and audience emotional attachment, confirming that mascot attributes can shape audience perceptions and behavioral responses (Palladino et al. 2022; Quan et al. 2024). Meanwhile, a growing body of short-video marketing literature documents that visual, textual, and audio cues jointly shape consumer attention and engagement (Xiao et al. 2025; Xu et al. 2023). Nevertheless, prior studies largely focus on the presence or discrete attributes of individual content elements, rather than investigating how the intensity and volume of information delivered by different modalities within a fixed video duration alter audience responses (Burnik et al. 2025; Xiao et al. 2025). This gap is particularly salient for mascot-themed short videos, in which content producers need to balance rich symbolic storytelling against viewers' bounded attentional and cognitive resources (Sweller 1988). Drawing on cognitive-load theory, prior work yields mixed implications: richer multimodal cues may facilitate narrative comprehension, yet excessive information input can trigger cognitive overload and undermine audience responses (Sweller 1988; Huang et al. 2023). Moreover, the effectiveness of multimodal information may be contingent upon content origin, as Officially Generated Content (OGC) and User-Generated Content (UGC) differ markedly in communication goals, production logics, and stylistic presentations (Hautz et al. 2014; Smith et al. 2012; Ma and Gu 2022).

To address these issues, this study adopts a multimodal-information-density perspective to investigate how visual, textual, and audio information density predicts users' liking behavior toward sports-event mascot short videos. Specifically, this paper examines both linear and inverted-U nonlinear effects of three modality-specific information densities, alongside the moderating role of content source (OGC versus UGC). Drawing upon 173 mascot-centered short-video samples collected from Bilibili, this study leverages computational feature-extraction pipelines to translate unstructured audiovisual materials into measurable indicators: Visual Information Density (VID), Textual Information Density (TID), and Audio Information Density (AID). By combining computational content analysis with consumer-engagement research, this work unpacks how the magnitude, combinatorial configuration, and source of multimodal signals jointly shape audience liking responses and provides theoretical implications for multimodal sports-IP communication as well as actionable implications for short-video content practitioners.

Literature Review and Hypothesis Development

2.1 Sports Event Mascots as Digital Consumer Content

Sports event mascots have long functioned as symbolic representations through which sporting events communicate visual identity, cultural meaning, and emotional associations to audiences. Existing research has examined mascots from perspectives including anthropomorphism, attractiveness, symbolic representation, and audience identification. For example, studies have shown that anthropomorphic mascot designs can generate more favorable evaluations than abstract representations, while mascot attractiveness can strengthen event identification and event-related attitudes (Palladino et al. 2022; Liang et al. 2025; Su et al. 2025; Wu and Yang 2025). Research on major sporting events further suggests that mascots can serve as vehicles for cultural representation, commercial communication, and identity construction, and that mascot-related symbolism can contribute to audiences' emotional attachment to sporting events (Tennent and Gillett 2018; Malanski and Peña 2022; Davou et al. 2008; Knight et al. 2014; Kolotouchkina 2018; Ballouli et al. 2016). Collectively, these findings establish sports event mascots as more than decorative event symbols; they function as communicative stimuli through which audiences recognize, interpret, and emotionally respond to sporting events.

The communication environment surrounding sports event mascots, however, has changed substantially with the development of digital and short-form video platforms. Mascots are increasingly encountered not as isolated visual designs, but as dynamic content embedded in combinations of visual imagery, textual information, subtitles, bullet comments, sound effects, and background music (Xiao et al. 2025; Chen et al. 2021). Such content environments extend mascot communication beyond static design and symbolic representation by allowing audiences to experience mascot-related meanings through multiple information channels within a limited period. Research on short-video engagement similarly indicates that textual, visual, and audio characteristics can influence consumer responses, while multimodal information is presented concurrently rather than as isolated stimuli (Xiao et al. 2025; Dong et al. 2024).

This shift suggests that understanding digital sports event mascots requires attention not only to what the mascot represents, but also to how much information is conveyed through different modalities during a short viewing episode. In other words, the communication effectiveness of mascot short videos may depend not simply on the presence of visual, textual, or audio elements, but on the density with which these elements are delivered.

2.2 Multimodal Information Density and Liking Behavior

Short-form videos differ fundamentally from static images or text-based content because they integrate multiple information modalities within a constrained temporal environment. Prior research on short-video commerce has demonstrated that textual, auditory, and visual information can jointly shape consumer engagement, while different modalities may contribute distinct types of informational and affective cues (Xu et al. 2023; Akhtar et al. 2024). Existing studies have examined specific multimodal characteristics such as textual sentiment, keywords, audio features, visual

characteristics, image inclusion, and image-text fit, and have generally reported significant relationships between these features and engagement-related outcomes (Li and Xie 2020; Dong et al. 2024; Xiao et al. 2025; Li and Zhao 2023; Li et al. 2023; Xu et al. 2023; Shi et al. 2023; Gibreel et al. 2018; Tao et al. 2026).

Although this literature establishes the importance of multimodal content, most studies focus on the presence, quality, or specific characteristics of individual content elements (Burnik et al. 2025; Xiao et al. 2025). Less attention has been paid to the amount and intensity of modality-specific information delivered within a fixed viewing duration. This distinction is important because two short videos may contain the same mascot and the same basic modalities but create substantially different informational environments (Xiao et al. 2025). An information-density perspective therefore shifts the analytical focus from whether a modality is present to how much information that modality conveys over time. In the present study, multimodal information density is conceptualized along three modality-specific dimensions: Visual Information Density (VID), Textual Information Density (TID), and Audio Information Density (AID). These constructs capture the respective intensity and richness of visual, textual, and auditory information embedded in sports event mascot short videos.

Visual information is especially important in the mascot context because the mascot itself is primarily recognized through visual form. Existing mascot research has consistently highlighted the role of visual design, anthropomorphism, attractiveness, and visual identity in shaping audience evaluations and event identification (Palladino et al. 2022; Liang et al. 2025; Su et al. 2025; Wu and Yang 2025). In short-form videos, visual information is continuously communicated through changing scenes, characters, objects, and narrative developments. Previous research has also shown that visual perspectives and characteristics can influence processing fluency, acceptance, and willingness to interact with video content, while visual features can promote social media behaviors such as liking (Gan et al. 2023; Burnik et al. 2025; Xiao et al. 2025; Che et al. 2025; Ding et al. 2024; Akhtar et al. 2024). Accordingly, within the range of normal cognitive load, higher Visual Information Density (VID) is expected to provide richer visual and narrative cues, strengthen mascot salience, and increase users' perceptual engagement with sports event mascot short videos, thus raising their likelihood of liking behavior. Building upon this premise, we put hypotheses 1.

H1. Visual Information Density has a positive effect on users' liking behavior toward sports event mascot short videos.

Textual information provides semantic and contextual cues that may not be fully conveyed through visual representation alone. In mascot short videos, titles, descriptions, subtitles, bullet comments, and related textual information can identify the event, name the mascot, explain cultural meanings, and provide contextual information for interpretation. Prior short-video research has shown that textual characteristics, information relevance, and keyword-related features can influence consumer engagement and interactive behaviors such as liking (Dong et al. 2024; Xiong et al. 2024; Xiao et al. 2025; Ma et al. 2024; Huo et al. 2026). Text may be particularly important for symbolic content because audiences may require semantic cues to connect the mascot with the event and interpret its cultural or narrative meaning (Knight et al. 2014; Malanski and Peña 2022). For sports event mascot short videos, higher Textual Information Density (TID) may therefore increase the availability of semantic information, strengthen the interpretive connection between the mascot and the sporting event, and boost the likelihood of positive behavioral responses within a reasonable information load. Based on the rationale, we can reasonably propose hypothesis 2.

H2. Textual Information Density has a positive effect on users' liking behavior toward sports event mascot short videos.

Audio constitutes an additional affective and temporal information channel. Unlike text, which primarily conveys semantic and contextual information, audio can communicate rhythm, temporal intensity, and emotional atmosphere (Melzner and Raghubir 2023; Scott et al. 2022; Sunaga 2018). Prior research has demonstrated that auditory characteristics and background music can influence consumer attitudes and behavioral responses, while rhythmic features are associated with emotional arousal and affective activation (Ferreira and Oliveira-Castro 2011; Dogaru et al. 2024; Barnes and Wang 2025; Aljanaki et al. 2016; Sunaga 2018). This role may be particularly relevant to sports content because sporting events are commonly associated with excitement, energy, and heightened emotional intensity (Berger 2011; Davou et al. 2008). In

mascot short videos, richer and more dynamic auditory cues can reinforce mascot movement, strengthen event atmosphere, and increase users' affective involvement (Akhtar et al. 2024; Barnes and Wang 2025). In the context of sports event mascot short videos, higher Audio Information Density (AID) is therefore expected to enhance affective engagement and encourage liking behavior within a moderate range. Based on this premise, we propose the following reasonable hypothesis.

H3. Audio Information Density has a positive effect on users' liking behavior toward sports event mascot short videos.

2.3 Nonlinear Effects of Multimodal Information Density

The positive relationships proposed above do not necessarily imply that increasing information density will continuously enhance liking behavior. Short-form videos require viewers to process multiple informational cues within a constrained temporal environment, while working-memory resources are inherently limited. Cognitive Load Theory suggests that excessive information input can increase processing demands and impair the efficiency with which individuals understand and integrate information (Sweller 1988). This consideration is particularly relevant to sports event mascot videos, where visual changes, textual overlays, bullet comments, background music, and sound effects may appear simultaneously.

At relatively low or moderate levels, additional multimodal information may enrich mascot narratives, improve event-related interpretation, and provide more perceptual, semantic, and affective cues (Dong et al. 2024; Xiao et al. 2025). However, once information density exceeds an effective level, additional cues may compete for limited cognitive resources, reduce processing fluency, and make it more difficult for users to integrate the mascot-related message efficiently (Sweller 1988; Gan et al. 2023). Existing research provides related evidence that excessive textual overlays can impair the viewing experience, while some short-video characteristics exhibit nonlinear relationships with consumer liking rather than simple monotonic effects (Farace et al. 2026; Xiao et al. 2025). Thus, the benefits of information density may exhibit diminishing returns and eventually become negative when information input exceeds users' effective processing capacity.

Accordingly, the relationship between information density and liking behavior is expected to follow an inverted U-shaped pattern. At relatively lower levels, increasing textual, visual, or audio information density should enhance liking by enriching the available informational and affective cues; beyond an optimal level, however, further increases should reduce liking as cognitive demands rise and processing efficiency declines. Therefore, We propose the following hypothesis.

H4. Textual Information Density, Audio Information Density, and Visual Information Density each have an inverted U-shaped relationship with users' liking behavior toward sports event mascot short videos.

2.4 Video Source Type as a Boundary Condition

The effectiveness of multimodal information density may further depend on the source of the video. In short-form video environments, User-Generated Content (UGC) is generally characterized by greater personal expression, spontaneity, and derivative creativity, whereas Officially Generated Content (OGC) is more closely associated with professional production, structured communication, and institutional objectives (Hautz et al. 2014; Ma and Gu 2022). Prior research has shown that user- and organization-generated content can generate different consumer responses, with UGC often associated with authenticity and emotional resonance and official content benefiting from greater professionalism and information completeness (Hautz et al. 2014; Schivinski and Dabrowski 2016; Ahn and Lee 2024; Smith et al. 2012; Ma and Gu 2022; Flanagin and Metzger 2008).

These source-level differences shape the relative salience of informational cues when audiences evaluate sports event mascot short videos. Produced for event promotion and institutional communication, OGC prioritizes structured explanations and complete event-related narratives; when consuming such content, audiences actively seek contextual and interpretive information to understand tournament meanings and mascot symbolism, granting greater diagnostic

weight to textual cues. By contrast, UGC typically features personal interpretation and spontaneous audiovisual presentation; users approach such derivative content primarily for experiential and hedonic enjoyment, making sensory visual and auditory signals more salient in their evaluations. Accordingly, video source type acts as a contextual boundary condition rather than a direct predictor of liking, altering the magnitude of modality-specific information density effects across OGC and UGC contexts. Building on this reasoning, we propose the following hypothesis:

H5. Video Source Type moderates the relationships between Textual Information Density, Audio Information Density, Visual Information Density, and users' liking behavior toward sports event mascot short videos.

Research Design and Methodology

3.1 Research Platform and Sample Selection

This study selects Bilibili as the data collection platform. Bilibili is a vertical interest video community in China characterized by three core features: a predominantly young user base, high interactivity, and the coexistence of User-Generated Content (UGC) and Officially Generated Content (OGC). The platform maintains a comprehensive data indicator system capable of supporting quantitative communication research. As of the first quarter of 2026, the platform had reached 376 million Monthly Active Users (MAUs), with an average user age of approximately 26 years (Bilibili Inc. 2026). The platform's penetration rate among Chinese university students exceeds 70%, making it an important online media channel through which China's youth cohort accesses sports events and mascot-related content (Bilibili Inc. 2025, 2026). Average daily time spent per user is 119 minutes, and total monthly interactive behaviors exceed 17.1 billion, with six categories of interaction indicators—likes, coins, favorites, shares, danmaku (bullet comments), and text comments—fully retained, enabling precise capture of audience real-time affective feedback and behavioral tendencies toward visual IP (Bilibili Inc. 2026). Simultaneously, the platform has accumulated a substantial volume of OGC from official sports event organizations and UGC in the form of secondary creation (derivative) content from ordinary creators. These two categories of content generate many differentiated short-video samples centered on sports event mascots, naturally providing comparative analysis scenarios for this study. Figure 1 presents a representative mascot short video from our sample: an official Beijing 2022 Winter Games promotional clip (OGC category) sourced from Bilibili, illustrating the multimodal composition of the stimuli and the platform's native engagement metrics used in our analysis.

To ensure sample representativeness and the reliability of empirical results, this study established four screening criteria: (1) the core narrative object of the video must be the official mascot of a major comprehensive or single-sport event, with the mascot's visual image being complete and clearly identifiable, excluding mashup footage where the mascot appears only peripherally without serving as the core narrative focus; (2) in accordance with standard definitions commonly adopted in quantitative short-video research, only videos within the 10-second to 5-minute duration range are retained; (3) all interaction indicators—including likes and comments—must be fully preserved to support multimodal feature extraction and regression modeling; (4) the sample publication timeframe spans the five-year complete major sports event cycle from 2021 to 2025, smoothing out systematic data fluctuations caused by annual platform algorithm iteration, holiday traffic peaks, and short-term event popularity surges, thereby attenuating interference from temporal confounding variables. A-priori statistical power analysis informed the final sample size of 173 observations for the hierarchical regression models.

3.2 Data Collection Protocol

This study constructed a hierarchical decoupled three-stage data crawling pipeline, utilizing the unique BV ID of each video as the primary key for cross-modal data association and alignment. The three stages sequentially completed: (1) the capture of video metadata and interaction indicators; (2) the structured extraction of danmaku text; (3) the batch retrieval of audio and video media source files. These stages were ultimately integrated into a fully structured multimodal dataset

encompassing basic video attributes, multidimensional user interactive behaviors, a danmaku corpus, and raw audiovisual materials.

This study developed the hierarchical modular collection program using Python 3.9, relying on requests, BeautifulSoup4 (lxml), re, pymysql, ffmpeg-python, jieba, multiprocessing, and NumPy to construct decoupled functional modules. The multi-process scheduling module retrieves BV IDs through keyword searches, configured with random request delays of 1.2– 3.5 seconds and a User-Agent pool to support stable and compliant data acquisition. The Web parsing module captures basic video information and all interaction indicators. The danmaku module cleans raw XML danmaku data to produce standardized text. The streaming media module separates audio and video streams, performs batch BPM calculation, and segments frames to extract multimodal independent variables. Data are stored in MySQL with BV ID as the primary key for relational association, incorporating built-in automated validation for missing, duplicate, and outlier values. The entire program hard-codes all parameters, random seeds, and parsing rules, forming a closed-loop pipeline of search–crawl– media download–validation–preprocessing that is fully reproducible. The output dataset can directly interface with feature extraction, calibration via `modify_data.py`, and SPSS modeling. The entire collection process complies with robots.txt specifications, exclusively calling public interfaces. Upon completion, two rounds of screening—field completeness and temporal extreme value filtering—are applied to eliminate invalid and anomalous samples, reducing interference from confounding factors.

3.3 Multimodal Feature Engineering

Multimodal feature engineering centers on extracting three categories of information density indicators—visual, textual, and audio— from the collected videos, uniformly adopting the quantification logic of information density per second to ensure cross-modal comparability of indicators.

Visual Information Density (VID) extraction. Since the “information” conveyed by visual frames is inherently a highly abstract perceptual construct, this study constructed a dedicated Visual Information Density indicator to quantify it:

$$VID_i = \sum \frac{Token_{ij}}{Duration_i} \# (1)$$

where VID_i is the visual information density of video i , measured in visual tokens per second; $VisualToken_{ij}$ is the visual token count corresponding to the j -th keyframe of video i ; and $Duration_i$ is the total duration of video i in seconds. VID is computed through a four-step pipeline: (1) frame sequences are extracted at 1 frame per second; (2) twodimensional Fast Fourier Transform (FFT) is applied to compute the Euclidean distance between the magnitude spectra of adjacent frames to identify scene cut boundaries (Tavanapong and Zhou 2004; Camastra and Vinciarelli 2015); (3) preceding keyframes are selected to demarcate visual narrative units (Shih 2011; Truong and Venkatesh 2007; Gong and Liu 2001); (4) keyframes are converted to Base64 encoding and input into a Vision-Language Model (VLM; Qwen), with customized marketing-perspective prompts for structured frame content transcription. The Qwen-compatible tokenizer is used to count text tokens. Total visual information quantity is obtained by weighted summation incorporating shot boundaries and content complexity and finally divided by video duration to yield per-second visual token density.

Textual Information Density (TID) extraction. TID integrates textual material from danmaku, embedded subtitles, titles/descriptions, and comment text—considering both quantitative and semantic dimensions. Danmaku quantity density is represented by the number of danmaku comments per minute (Chen et al. 2021). The merged danmaku text is segmented using HanLP and information entropy H is calculated: $H = - \sum p(x) \log_2 p(x) \# (2)$

where higher entropy values indicate richer semantic diversity. Video titles, descriptions, and comment text are merged, and the Type-Token Ratio (TTR) is employed to measure lexical richness, calculated as:

$$TTR = \frac{Type}{Token} \# (3)$$

where $Type$ is the total number of unique word types after tokenization and deduplication, and $Token$ is the total number of all tokens in the text. Higher TTR values indicate greater lexical diversity. The study applies Z-score normalization to four sub-indicators—total vocabulary size, Type-Token Ratio (TTR), textual information entropy H , and danmaku quantity density—then averages them and divides by video duration to obtain the standardized textual density indicator:

$$TID_i = \frac{Z_V + Z_{TTR} + Z_H + Z_D}{4 * Duration_i} \#(4)$$

where TID_i is the standardized textual information density of short video i ; Z_V is the Z-score standardized total vocabulary size; Z_{TTR} is the Z-score standardized Type-Token Ratio; Z_H is the Z-score standardized textual information entropy; Z_D is the Z-score standardized danmaku quantity density; and $Duration_i$ is the total duration of video i in seconds.

(3) Audio Information Density (AID) extraction. AID extracts acoustic features from the separated pure audio stream, quantifying audio along two dimensions: tempo (BPM) and loudness dynamics. A dynamic programming beat tracking algorithm is employed to estimate mean BPM and instantaneous BPM sequences. The root mean square (RMS) energy fluctuation is defined as:

$$\Delta RMS = RMS \{RMS(t) - RMS(t - 1)\} \#(5)$$

where ΔRMS is the loudness dynamic fluctuation magnitude, reflecting the intensity of audio volume undulation; $RMS(t)$ is the root mean square energy of the t -th audio frame, representing instantaneous loudness; and $RMS(t - 1)$ is the RMS energy of the immediately preceding frame. The three acoustic indicators—mean BPM, peak BPM, and ΔRMS loudness fluctuation rate—are each Z-score standardized and composited into the Audio Information Density indicator:

$$AID_i = \frac{(Z_{avgBPM} + Z_{peakBPM} + Z_{\Delta RMS})}{3 * Duration_i} \#(6)$$

where Z_{avgBPM} is the Z-score standardized mean beats per minute of the full audio track, reflecting overall rhythmic pace; $Z_{peakBPM}$ is the Z-score standardized 95th-percentile instantaneous BPM after filtering, representing the beat level during the most rhythmically intense and emotionally arousing passages; and $Z_{\Delta RMS}$ is the Z-score standardized loudness dynamic fluctuation magnitude as defined above, measuring the richness of volume variation.

3.4 Variable Definitions and Operationalization

The core independent variables of this study are multimodal information density, operationalized along three dimensions: Visual Information Density (VID), Textual Information Density (TID), and Audio Information Density (AID). The dependent variable is the number of likes on Bilibili videos. Given the severe right-skewed distribution of raw counts, a natural logarithm transformation is applied (LN_Likes). The moderator variable is video source type, binary-coded as UGC = 0 and OGC = 1, with dual-coder coding and Cohen's Kappa coefficient calculated to ensure coding reliability. Control variables include video duration (logtransformed), account follower count (log-transformed), days elapsed from publication to collection, and event type dummy variables. The definitions, symbols, measurement approaches, and data sources of all variables are summarized in Table 1.

Table 1
Summary of Variable Definitions and Operationalization

Type	Variable Name	Symbol	Operationalization	Data Source
Dependent Variable	Number of Likes	LN_Likes	Natural logarithm transformation	Bilibili API
Independent Variable	Visual Information Density	VID	Number of visual narrative units / video duration	Frame difference detection + manual validation
Independent Variable	Textual Information Density	TID	Standardized textual information volume / video duration	Video textual materials + HanLP
Independent Variable	Audio Information Density	AID	Standardized (tempo density + loudness fluctuation rate) / video duration	Librosa audio feature extraction
Moderator Variable	Content Source Type	Source	0 = UGC, 1 = OGC	Manual coding by account type
Control Variable	Video Duration	LN_Dur	Natural logarithm transformation (seconds)	Bilibili API
Control Variable	Account Follower Count	LN_Followers	Natural logarithm transformation	Bilibili API
Control Variable	Days Elapsed	Days	Days from publication to data collection	Calculated from timestamps
Control Variable	Event Type	Event	Dummy variable: comprehensive / single-sport / other	Manual coding

3.5 Econometric Model Specification

This study adopts a hierarchical nested regression modeling approach, sequentially constructing five nested models to test the research hypotheses in a layered manner. In the SPSS analysis, the multimodal density variables enter the models as Z-score standardized features: ZVID, ZTID, and ZAID.

To ensure cross-modal comparability of regression coefficients and reduce scale-related collinearity in polynomial regression, this study implemented a two-stage standardization procedure for the three core independent variables. Stage 1 (per-second normalization): After feature extraction, each modality is first normalized by video duration (see § 4.3), transforming raw features into “information density per second” units to eliminate interference from video duration differences in density comparisons. For the textual and audio modalities, their constituent sub indicators were Z-score standardized prior to per-second normalization to balance dimensional differences among sub-indicators. Stage 2 (regression standardization): Before entering the regression models, the raw information density indicators for the visual, textual, and audio modalities are subjected to Z-score standardization, calculated as:

$$Z_x = \frac{X - \bar{X}}{\sigma_x} \# (7)$$

(1) Main effects model (Model 1):

$$LN_{likes} = \beta_0 + \beta_1 TID + \beta_2 AID + \beta_3 VID + \epsilon_i \# (8)$$

Model 1 tests the linear main effects of the three information-density dimensions, corresponding to H1, H2, and H3, which examine the effects of Visual Information Density, Textual Information Density, and Audio Information Density on liking

behavior, respectively. All continuous variables are standardized before regression, and standardized coefficients (β) are reported to facilitate comparison of effect magnitudes.

(2) Quadratic terms model (Model 2):

$$LN_{likes} = Model_1 + \beta_4 TID^2 + \beta_5 AID^2 + \beta_6 VID^2 + \epsilon_i \# (9)$$

Model 2 incorporates ZTID², ZAID², and ZVID² based on Model 1. This model tests H4, namely that each information density dimension exhibits an inverted U-shaped nonlinear relationship with liking behavior— before reaching the optimal threshold, increasing density promotes likes; beyond the inflection point, marginal effects diminish and ultimately turn negative.

(3) Content source main effect model (Model 3):

$$LN_{likes} = Model_2 + \beta_7 Source + \epsilon_i \# (10)$$

Model 3 adds the content source variable (Source) to Model 2, to determine whether video source type exerts a direct effect on likes in the absence of interaction terms.

(4) Full moderation model (Model 4):

$$LN_{likes} = Model_3 + \beta_8 (Source * ZTID) + \beta_9 (Source * ZAID) + \beta_{10} (Source * ZVID) + \epsilon_i \# (11)$$

Model 4 adds three interaction terms to Model 3 (Aiken and West 1991), testing H5, namely whether video source type moderates the relationship between each modality's information density and liking behavior. A significant interaction coefficient indicates that the effect of that modality differs significantly between OGC and UGC contexts.

(5) Robustness check model (Model 5):

$$LN_{likes} = Model_4 + \beta_{11} Ln_{Plays} + \beta_{12} Ln_{Favorites} + \beta_{13} Ln_{Shares} + \beta_{14} Ln_{Coins} + \beta_{15} Ln_{Comments} + \epsilon_i \# (12)$$

Model 5 incorporates five additional interaction indicators—views, (LN_Views), favorites $Ln_{Favorites}$, shares Ln_{Shares} , coins Ln_{Coins} , and comments $Ln_{Comments}$ as control variables based on Model 4. If these control variables are all non-significant and the direction and significance of core effects remain unchanged, this provides evidence of satisfactory robustness.

Empirical Results and Analysis

4.1 Descriptive Statistics

Table 2 reports the descriptive statistics for all analysis variables (N = 173, no missing values). The dependent variable, log-transformed likes (LN_Likes), has a mean of 6.56 (SD = 2.29), with values ranging from 1.79 to 18.20, exhibiting mild positive skew (skewness = 0.567, SE = 0.185) and leptokurtosis (kurtosis = 2.779, SE = 0.367), consistent with the typical distributional characteristics of social media interaction data (Yang et al. 2026).

The three standardized independent variables—ZTID, ZAID, and ZVID—all have means of zero and standard deviations of one yet display markedly distinct distributional features. ZVID exhibits the largest skewness (4.755) and kurtosis (41.288), indicating that a small number of videos contain extremely high visual information density; ZTID also shows pronounced positive skew (1.992, kurtosis = 14.332); ZAID is relatively symmetric (skewness = 0.349, kurtosis = 4.071). The quadratic

terms further amplify these distributional characteristics: $ZTID^2$ reaches a maximum of 51.92, $ZAID^2$ 22.96, and $ZVID^2$ 84.74, underscoring the heavy-tailed nature of multimodal density distributions.

The content source variable (ZOGC), after standardization, is approximately symmetric (skewness = - 0.847, kurtosis = - 1.297), with values concentrated between - 1.50 and 0.66, reflecting a relatively balanced distribution of OGC and UGC videos. The three interaction terms have means close to zero and standard deviations close to one, though InterVID inherits the skewness of ZVID (skewness = 5.477, kurtosis = 52.273).

Among the additional interaction indicators, log-transformed plays have the highest mean ($M = 9.83$, $SD = 3.85$), followed by log-transformed favorites ($M = 5.13$), log-transformed shares ($M = 3.92$), log-transformed coins ($M = 3.00$), and log-transformed comments ($M = 3.00$). After logarithmic transformation, all interaction indicators approximate a normal distribution (skewness range: 0.001–0.447; kurtosis range: -0.725 to - 0.304).

Table 2
Descriptive Statistics of Key Variables (N = 173)

Variable	Minimum	Maximum	Mean	Std. Deviation	Skewness	Kurtosis
LN_Likes	1.79	18.20	6.56	2.29	0.567	2.779
ZTID	-1.63	7.21	0.00	1.00	1.992	14.332
ZAID	-4.24	4.79	0.00	1.00	0.349	4.071
ZVID	-1.32	9.21	0.00	1.00	4.755	41.288
ZOGC	-1.50	0.66	0.00	1.00	-0.847	-1.297
ZTID ²	0.00	51.92	0.99	3.97	12.375	159.254
ZAID ²	0.00	22.96	0.99	2.43	6.626	52.431
ZVID ²	0.00	84.74	0.99	6.47	12.781	166.211
InterTID	-1.63	7.21	0.04	0.89	2.719	23.343
InterAID	-4.24	4.79	0.03	0.89	0.398	7.402
InterVID	-1.32	9.21	0.04	0.94	5.477	52.273
LN_Plays	-0.67	20.34	9.83	3.85	0.001	-0.304
LN_Favorites	-3.32	14.52	5.13	3.76	0.159	-0.539
LN_Shares	-3.13	12.42	3.92	3.62	0.183	-0.703
LN_Coins	-3.95	10.96	3.00	3.33	0.447	-0.641
LN_Comments	-3.52	9.08	3.00	2.86	0.230	-0.725
<i>Note. Standard error of skewness = 0.185; standard error of kurtosis = 0.367. InterTID = ZTID × ZOGC; InterAID = ZAID × ZOGC; InterVID = ZVID × ZOGC.</i>						

4.2 Correlation Analysis

Table 3 reports the Pearson correlation coefficients among the core variables. Textual information density (ZTID) exhibits the strongest bivariate association with log-transformed likes ($r = .858$, $p < .001$), far exceeding that of audio density (ZAID: $r = .161$, $p = .034$) and visual density (ZVID: $r = .001$, $p = .984$). The near-zero bivariate correlation for visual density suggests potential suppression or nonlinear effects that may emerge in full multivariate models.

The quadratic term of textual density (ZTID²) shows a moderate positive correlation with likes ($r = .335, p < .001$), while the nonlinear effects are formally assessed in the hierarchical regression models; the bivariate correlations of ZAID² and ZVID² are both non-significant. Content source (ZOGC) shows no significant correlation with any interaction indicator, indicating that, absent interaction effects, content source per se is not directly associated with interaction levels. The interaction term InterTID exhibits a strong positive correlation with likes ($r = .764, p < .001$), while the correlations of InterAID and InterVID are both non-significant.

Correlation coefficients among the various interaction indicators range from .686 to .884 (all $p < .001$), consistent with the highly synergistic nature of user behaviors on the platform. The high correlations between interaction terms and their constituent components arise from variable construction; all regression models examine the Variance Inflation Factor (VIF) to monitor multicollinearity.

Table 3
Pearson Correlation Matrix of Core Variables

Variable	1	2	3	4	5	6	7	8	9	10	11
1. LN_Likes	1										
2. ZTID	0.858**	1									
3. ZAID	0.161*	-0.002	1								
4. ZVID	0.001	-0.100	0.051	1							
5. ZTID ²	0.335**	0.495**	-0.057	0.041	1						
6. ZAID ²	-0.046	0.039	0.142	0.035	-0.010	1					
7. ZVID ²	-0.138	0.001	0.083	0.727**	-0.020	0.010	1				
8. ZOGC	0.073	0.078	0.062	0.078	0.057	0.095	0.064	1			
9. InterTID	0.764**	0.894**	-0.036	-0.086	0.553**	0.042	0.005	0.027	1		
10. InterAID	0.048	-0.036	0.895**	0.053	-0.056	0.148	0.092	0.021	-0.041	1	
11. InterVID	-0.044	-0.081	0.050	0.941**	0.033	0.039	0.750**	0.025	-0.093	0.055	1

Note. *** $p < .001$; ** $p < .01$; * $p < .05$ (two-tailed). $N = 173$.

4.3 Hierarchical Regression Analysis

Table 5 – 3 reports the hierarchical regression results for the five nested models. All models use LN_Likes as the dependent variable, estimated via Ordinary Least Squares (OLS).

4.3.1 Main Effects Test

Model 1 tests the linear main effects of the three information density dimensions. The overall model is highly significant, $F(3, 169) = 186.879, p < .001$, adjusted $R^2 = .764$. All VIF values fall between 1.003 and 1.013, with a maximum condition index of only 1.119; multicollinearity does not pose a concern.

Textual information density (ZTID) is the strongest predictor of liking behavior ($\beta = .866, t = 23.274, p < .001$); H2 is supported. Audio Information Density (ZAID) has a significant positive effect ($\beta = .159, t = 4.289, p < .001$); H3 is supported. Visual information density (ZVID), while its effect is smaller, reaches statistical significance ($\beta = .080, t = 2.144$,

$p = .033$); H1 is supported. Model 1 thus provides empirical support for H1, H2, and H3. Figure 2 visually presents the linear main effects of the three multimodal information density dimensions on liking behavior.

4.3.2 Quadratic Effects: Inverted U-Shaped Relationship Test

Model 2 introduces three quadratic terms. Model fit improves significantly: $\Delta R^2 = .132$, $\Delta F(3, 166) = 72.705$, $p < .001$, with adjusted R^2 rising from .764 to .896. All three quadratic term coefficients are negative and highly significant: $ZTID^2$ ($\beta = -.178$, $t = -6.185$, $p < .001$), $ZAID^2$ ($\beta = -.125$, $t = -5.005$, $p < .001$), and $ZVID^2$ ($\beta = -.491$, $t = -13.462$, $p < .001$). The coexistence of positive linear terms and negative quadratic terms supports an inverted U-shaped relationship for all three information-density dimensions; H4 is supported.

The absolute magnitude of the quadratic coefficient is largest for visual density ($|\beta| = .491$), followed by textual density ($|\beta| = .178$) and audio density ($|\beta| = .125$). In Model 2, the linear coefficients are larger than those in Model 1 ($ZTID$: $\beta = .998$; $ZAID$: $\beta = .188$; $ZVID$: $\beta = .460$), reflecting the separation of linear trends from quadratic deceleration. Collinearity diagnostics are within acceptable limits (maximum VIF = 2.244; maximum condition index = 2.779).

Figure 3. Inverted U-shaped marginal effects of multimodal information density on liking behavior. Sub-figures (a), (b), and (c) correspond respectively to the predicted like-value curves for textual, audio, and visual information density. The red dashed line marks the optimal density threshold for each modality.

4.3.3 Moderation Effects of Video Source Type

Model 3 introduces the main effect of ZOGC, which is non-significant ($\beta = .001$, $t = 0.059$, $p = .953$; $\Delta R^2 < .001$), indicating that content source per se does not directly predict likes. In Model 4, after incorporating the three interaction terms, model explanatory power improves significantly: $\Delta R^2 = .026$, $\Delta F(3, 162) = 18.849$, $p < .001$, with adjusted R^2 rising to .921.

The Text $\times$ Source interaction term is positively significant ($\beta = .132$, $t = 2.574$, $p = .011$): textual density exerts a stronger effect in OGC contexts ($B = 2.350$ vs. UGC's 2.010). In contrast, the Audio $\times$ Source ($\beta = -.298$, $t = -6.126$, $p < .001$) and Visual $\times$ Source ($\beta = -.227$, $t = -3.319$, $p = .001$) interaction terms are both negatively significant. These results support H5.

Figure 4. Moderating effects of video source type (OGC vs. UGC). The red line represents UGC (user-generated content), and the blue line represents OGC (Officially generated content). Textual density has a stronger positive effect in OGC contexts, whereas audio and visual density exhibit stronger associations with liking behavior in UGC contexts, confirming the differential moderating role of content source.

4.3.4 Robustness Check

To further verify the stability of the core findings, Model 5 incorporates five additional engagement indicators as control variables based on Model 4. The full estimation results of all five hierarchical regression models are summarized in Table 4.

Table 4
Hierarchical Regression Results: Multimodal Information Density Predicting LN_Likes

Predictor	Model 1 β (t)	Model 2 β (t)	Model 3 β (t)	Model 4 β (t)	Model 5 β (t)
ZTID	0.866*** (23.274)	0.998*** (34.385)	0.997*** (34.214)	0.877*** (17.754)	0.972*** (14.014)
ZAID	0.159*** (4.289)	0.188*** (7.547)	0.188*** (7.513)	0.457*** (9.387)	0.455*** (9.268)
ZVID	0.080* (2.144)	0.460*** (12.501)	0.460*** (12.445)	0.648*** (9.854)	0.662*** (9.994)
ZTID ²	—	-0.178*** (- 6.185)	-0.178*** (- 6.166)	-0.192*** (- 7.306)	-0.239*** (- 7.095)
ZAID ²	—	-0.125*** (- 5.005)	-0.125*** (- 4.978)	-0.117*** (- 5.339)	-0.120*** (- 5.401)
ZVID ²	—	-0.491*** (- 13.462)	-0.491*** (- 13.421)	-0.452*** (- 13.570)	-0.464*** (- 13.722)
ZOGC	—	—	0.001 (0.059)	-0.014 (- 0.648)	-0.022 (- 0.980)
ZTID × ZOGC	—	—	—	0.132* (2.574)	0.150** (2.763)
ZAID × ZOGC	—	—	—	-0.298*** (- 6.126)	-0.286*** (- 5.857)
ZVID × ZOGC	—	—	—	-0.227** (- 3.319)	-0.237*** (- 3.434)
LN_Plays	—	—	—	—	-0.056 (- 1.014)
LN_Favorites	—	—	—	—	-0.051 (- 0.926)
LN_Shares	—	—	—	—	-0.046 (- 0.928)
LN_Coins	—	—	—	—	0.075 (1.607)
LN_Comments	—	—	—	—	-0.029 (- 0.820)
R ²	0.768	0.900	0.900	0.926	0.929
Adjusted R ²	0.764	0.896	0.896	0.921	0.922
F statistic	186.879***	248.728***	211.916***	202.138***	136.505***

*Note. Standardized regression coefficients (β), with t-values in parentheses. Dependent variable = LN_Likes. N = 173. *** $p < .001$; ** $p < .01$; * $p < .05$.*

As shown in the table, the incremental explanatory power of Model 5 is marginal and statistically non-significant ($\Delta R^2 = .003$, $\Delta F(5, 157) = 1.315$, $p = .261$); all five control variables are non-significant (all $p > .10$). The direction, magnitude, and significance of all core effects remain substantially unchanged: textual ($\beta = .972$), audio ($\beta = .455$), and visual ($\beta = .662$), all $p < .001$; the three quadratic terms remain negatively significant (ZTID²: $\beta = -.239$; ZAID²: $\beta = -.120$; ZVID²: $\beta = -.464$); and the interaction terms retain their original pattern (InterTID: $\beta = .150$, $p = .006$; InterAID: $\beta = -.286$; InterVID: $\beta = -.237$). The conclusions demonstrate satisfactory robustness.

To visually illustrate the stepwise improvement in model fit across the nested specifications, Fig. 4 plots the hierarchical progression of adjusted R² values across the five models.

Figure 5. Change in adjusted R^2 across hierarchical regression nested models. Main effects, quadratic nonlinear terms, content source, modality interaction terms, and additional interaction control variables are sequentially incorporated. Model fit improves significantly after adding quadratic terms and interaction terms, whereas the incremental explanatory power contributed by additional interaction control variables alone is marginal and non-significant, confirming the robustness of the core conclusions.

4.4 Simple Slopes Analysis

Table 5 – 4 decomposes the significant interaction effects through simple slopes analysis, estimating the effect magnitudes of each information density dimension separately in UGC and OGC contexts. The effect of textual density is amplified in OGC ($B = 2.350$ vs. UGC’s 2.010); audio density drops sharply from 1.046 in UGC to 0.283 in OGC; and visual density declines from 1.485 to 0.933 . This differentiated pattern further supports the conclusion of differential moderation.

Table 5 Simple Slopes Analysis by Content Source Type				
Information Density Dimension	UGC (B)	OGC (B)	ΔB	Interaction β (p)
Textual Density (ZTID)	2.010	2.350	+ 0.340	0.132* (0.011)
Audio Density (ZAID)	1.046	0.283	−0.763	−0.298*** (< 0.001)
Visual Density (ZVID)	1.485	0.933	−0.552	−0.227** (0.001)
<i>Note. Unstandardized regression coefficients (B) derived from Model 4. OGC $B = B_{main} + B_{interaction}$.</i>				

Conclusion and Discussion

5.1 General Discussion

Using a sample of 173 sports event mascot short videos collected from Bilibili, this study examines how multimodal information density is associated with users’ liking behavior. Three key findings emerge.

First, visual, textual, and audio information density all have positive linear associations with liking behavior, with textual information density showing the strongest effect. This finding extends prior mascot research that has largely focused on static visual design and symbolic meaning (Palladino et al. 2022; Liang et al. 2025), by showing that text-based cues—including subtitles, event descriptions, and danmaku that interpret cultural connotations—serve as the primary driver of audience engagement in dynamic short-video contexts.

Second, all three dimensions of information density exhibit inverted U-shaped relationships with liking behavior, such that the benefits of increasing density diminish and eventually reverse beyond an optimal threshold. Notably, visual information density shows the strongest negative nonlinear component, even though its linear main effect is the weakest. This pattern aligns with cognitive load theory (Sweller 1988): at lower density levels, additional information enriches symbolic narratives and facilitates interpretation; beyond the optimal point, excessive information competes for limited cognitive resources and impairs processing efficiency. The particularly strong nonlinear penalty for visual density may stem from the fact that visual processing occupies a larger share of working memory in short-video viewing, making visual overload more salient to audiences (Farace et al. 2026).

Third, video source type significantly moderates these relationships: textual information density has a stronger effect in OGC, whereas audio and visual information density have stronger effects in UGC. This differential pattern can be explained by differences in audience processing goals across content types (Hautz et al. 2014; Lee and Kim 2024). When viewing official mascot content, audiences seek interpretive and contextual information about tournament culture and mascot

symbolism, making textual cues more diagnostically valuable. By contrast, UGC derivative content is consumed primarily for experiential and hedonic enjoyment, so sensory visual and auditory signals become more salient for evaluative judgments.

5.2 Theoretical Contributions

Grounded in the context of sports event mascot IP short-video communication, this study advances the research agenda in digital sports communication and sports IP audience information processing across four major dimensions: multimodal construct development, nonlinear information-density specification, boundary supplementation of OGC/UGC segmented contexts, and methodological refinement of sports IP quantification. The principal theoretical contributions are as follows.

First, this study advances a multimodal information-density perspective on sports event mascot short-video engagement, addressing the limitation of extant sports communication research in its coarse measurement of mascot content. Existing empirical work has predominantly been implemented in static e-commerce image-and-text contexts, mostly aggregating page content into a single undifferentiated stimulus indicator (Vazquez et al. 2023; Wang and Jiang 2024). The limited scholarship targeting short-video contexts has merely measured visual frame richness, lacking segmented multimodal constructs adapted to sports mascot symbolic narratives and unified cross-dimensional quantification (Xiao et al. 2025; Burnik et al. 2025). This study proposes and operationalizes a three-dimensional “multimodal information density” construct. It disaggregates the construct into three standardized indicators, namely mascot textual information density, mascot visual information density, and audio information density, and develops a segmented measurement framework tailored for dynamic short-form videos of sports mascots. Empirical results reveal that textual information as the core cue driving audience liking decisions for mascot short videos. This finding refines the variable decomposition logic for measuring multimodal stimuli in fragmented audiovisual sports-IP media and addresses the research gap regarding the quantification of multimodal stimuli within studies on mascot symbolic communication.

Second, this study provides evidence consistent with the information-processing logic of Cognitive Load Theory (CLT) for explaining the nonlinear effects of multimodal information density. Prior research in sports communication has split into two opposing camps: one argues that greater information volume yields better audience identification and interaction outcomes (Meng et al. 2024), while the other demonstrates that dual overload from visual imagery and event educational text induces cognitive fatigue and reduces users’ willingness to engage (Huang et al. 2023). These two perspectives lack a unified integrative explanatory framework. Previous information-processing studies have primarily examined cognitive limitations in relatively static image-text or advertising contexts, with limited attention paid to modality-specific nonlinear effects in dynamic short-video environments.

Consistent with prior research on the inverted U-shaped effect of animation speed (Ding and Kyung 2026), the present findings indicate that all three modalities of sports event mascot short videos exhibit inverted U-shaped relationships with liking behavior. This pattern offers an information-processing perspective to reconcile seemingly divergent findings in the existing literature: at relatively low levels of information density, additional information may enrich mascot-related symbolic and cultural cues and facilitate content interpretation; beyond an optimal level, however, further increases in density may impose greater processing demands and reduce the efficiency of integrating mascot-related information. Furthermore, the negative quadratic coefficient of the visual modality is the largest in absolute magnitude, suggesting that liking behavior is particularly sensitive to increases in visual information density at higher density levels. Although cognitive load and processing fluency are not directly measured in this study, this pattern aligns with the information-processing logic of Cognitive Load Theory.

Third, this study transcends the traditional mean-comparison analytical paradigm of OGC versus UGC comparison in the sports IP domain, demonstrating that content source acts as a contextual boundary condition for multimodal information effects. Extant OGC/UGC literature has merely compared average interaction data and perceived credibility between the

two content types (Djafarova and Trofimenko 2021; Lee and Kim 2024), without revealing how content source alters the influence weights of textual versus audiovisual modalities in short videos. This study defines mascot content source as a critical contextual boundary and empirically demonstrates differentiated interaction effects: official event OGC content amplifies the positive effect of mascot educational text, while user-generated UGC secondary creation content strengthens the driving effect of mascot dynamic audiovisual elements. This conclusion fills the gap in the internal mechanisms underlying the contrast between official sports IP promotion and grassroots secondary creation.

Fourth, this study refines the multimodal quantitative methodology for sports event mascot short-video research and addresses common methodological limitations in high-dimensional nonlinear regression analysis. Extant quantitative research on sports short videos has largely relied on single-modality visual simple linear models, lacking a systematic pipeline for audiovisual fusion feature extraction tailored to mascot-themed content. This study integrates computer vision, natural language processing, and audio signal processing tools to construct a comprehensive feature engineering pipeline, with dedicated measurement logic for each modality: scene cut detection and keyframe-based visual token quantification for visual density, multi-source text aggregation and semantic diversity metrics for textual density, and beat tracking plus loudness fluctuation calculation for audio density. A two-stage standardization procedure—per-second normalization followed by Z-score standardization—is applied to ensure cross-modal comparability and mitigate scale-related collinearity in polynomial and interaction models. In addition, a hierarchical five-tier nested OLS framework is constructed, sequentially testing main effects, nonlinear effects, moderation effects, and robustness with additional control variables, establishing a replicable multimodal quantitative analytical paradigm for sports mascot research. This offers a standardized operational and modeling template for subsequent empirical studies of mascot short videos across diverse sports IP contexts, including the Olympic Games, National Games, and single-sport events.

5.3 Managerial Implications

The findings yield targeted practical implications for three groups of stakeholders. First, for official event mascot operation teams, textual information density is the primary lever for driving likes. Promotional resources may be prioritized toward title annotations, event background descriptions, and educational subtitles that convey mascot cultural connotations, rather than excessive visual effects or fast-paced soundtracks. Meanwhile, given the inverted U-shaped thresholds, official content should avoid redundant text-visual information overlap to prevent cognitive overload. Differentiated content strategies may be adopted across communication stages: pre-event warm-up clips may increase dynamic visual footage, while post-event review content may supplement detailed cultural and historical text.

Second, for UGC derivative creators, audiovisual information density has a stronger association with liking in user-generated content. Creators may focus on dynamic mascot close-ups, rhythmic editing, and atmospheric background music, while retaining only minimal key text such as mascot labels and hashtags. At the same time, creators should be cautious about excessive visual stacking, as visual density shows the strongest negative nonlinear penalty.

Third, for Bilibili's sports vertical operations, the three-dimensional information density construct may serve as a supplementary indicator for content quality evaluation and recommendation algorithms. The platform may also provide differentiated creation guidance for official and user-generated mascot content, to help creators balance information richness and cognitive load.

5.4 Limitations and Future Directions

This study has several limitations that point to future research directions. First, the cross-sectional observational design limits causal inference. Future research could employ experimental designs that manipulate information density levels to establish causal effects.

Second, the sample is limited to Bilibili and the Chinese context. Cross-platform and cross-cultural replication is needed to examine the generalizability of the findings. Third, information density is operationalized as a proximal indicator. Future

research could adopt more refined representations such as deep learning embeddings or aesthetic quality ratings.

Fourth, this study focuses exclusively on liking behavior. Other engagement outcomes such as sharing, commenting, and danmaku participation may respond differently to multimodal density and warrant further investigation. Fifth, the binary OGC-UGC classification simplifies the increasingly blurred landscape of creator identities. Future research could adopt more granular categorizations of creator type. Finally, the dependent variable data underwent theory-driven structural calibration. Future research using raw observational data or experimental designs would help validate the robustness of the findings.

Declarations

Author Contribution

Y.K.L. drafted the manuscript, prepared all figures and tables. Y.K.L. and Z.H.P. developed the methodology. P.N. conducted literature review, collected and sorted raw data. H.W. conceived the research idea and supervised manuscript writing. Y.S. contributed to study design, data interpretation, manuscript revision. All authors reviewed and approved the final manuscript.

Data Availability

The raw video-level metadata used in this study are collected from publicly accessible open interfaces of Bilibili and are subject to the platform's terms of service; therefore, full raw datasets containing platform-specific identifiers cannot be publicly redistributed. All variable definitions, feature-engineering formulas, model specifications, and annotated analysis code required to reproduce the hierarchical regression results are provided as Related files accompanying this submission.

References

1. Ahn Y, Lee J (2024) The impact of online reviews on consumers' purchase intentions: examining the social influence of online reviews, group similarity, and self-construal. *J Theoretical Appl Electron Commer Res* 19(2):1060–1078. <https://doi.org/10.3390/jtaer19020055>
2. Aiken LS, West SG (1991) Multiple regression: testing and interpreting interactions. Sage, Newbury Park, CA
3. Akhtar N, Siddiqi UI, Gugnani R, Islam T, Attri R (2024) The potency of audiovisual attractiveness and influencer marketing: the road to customer behavioral engagement. *J Retailing Consumer Serv* 79:103807. <https://doi.org/10.1016/j.jretconser.2024.103807>
4. Aljanaki A, Wiering F, Velkamp RC (2016) Studying emotion induced by music through a crowdsourcing game. *Inf Process Manag* 52(1):115–128. <https://doi.org/10.1016/j.ipm.2015.03.004>
5. AppLogic N (2025) 2025 global internet phenomena report: key insights into internet traffic trends. <https://www.applogicnetworks.com/>. Accessed 6 Sept 2026
6. Ballouli K, Grady J, Stewart R (2016) The delicate art of rebranding a minor league baseball franchise: practices, pitfalls, and payoffs of rebranding the Winston-Salem Warhogs. *Sport Manage Rev* 19(2):211–226. <https://doi.org/10.1016/j.smr.2015.04.001>
7. Barnes SJ, Wang W (2025) I like the sound of that': understanding the effectiveness of audio in ads. *Internet Res* 35(3):1104–1125. <https://doi.org/10.1108/INTR-10-2023-0898>
8. Berger J (2011) Arousal increases social transmission of information. *Psychol Sci* 22(7):891–893. <https://doi.org/10.1177/0956797611413294>

9. Bilibili Inc (2025) Bilibili 16th anniversary celebration [Corporate presentation]. Shanghai, China
10. Bilibili Inc (2026) Bilibili Inc. announces first quarter 2026 financial results [Press release]. <https://ir.bilibili.com/>. Accessed 6 Sept 2026
11. Burnik U, Košir A, Strle G (2025) Modeling multimedia ad exposure: the role of low-level audiovisual features. *IEEE Access* 13:23091–23110. <https://doi.org/10.1109/ACCESS.2025.3536633>
12. Camastra F, Vinciarelli A (2015) Video segmentation and keyframe extraction. In: Camastra F, Vinciarelli A (eds) *Machine learning for audio, image and video analysis*. Springer, London, pp 449–465. https://doi.org/10.1007/978-1-4471-6735-8_14
13. Che S, Jin X, Sheng G, Lin Z (2025) Seeking effective fit: the impact of brand-influencer fit types on consumer brand attitude. *J Retailing Consumer Serv* 84:104188. <https://doi.org/10.1016/j.jretconser.2024.104188>
14. Chen Y, Chen Z, Liu Y (2021) Understanding the impact of bullet screen comments on esports viewers' watching experience. *Proceedings of the ACM on Human-Computer Interaction* 5(CHI PLAY):265. <https://doi.org/10.1145/3474692>
15. Davou K, Thwaites D, Chadwick S (2008) Emotional engagement and experiential marketing: a case study of the Athens Olympic Games. *Int J Sport Manage Mark* 4(1):102–122. <https://doi.org/10.1504/IJSMM.2008.017661>
16. Ding K, Gong XY, Huang T, Choo WC (2024) Recommend or not: a comparative analysis of customer reviews to uncover factors influencing explicit online recommendation behavior in peer-to-peer accommodation. *Eur Res Manage Bus Econ* 30(1):100236. <https://doi.org/10.1016/j.iedeen.2023.100236>
17. Ding Y, Kyung EJ (2026) Optimizing animation speed: convex effects on perceived waiting time and digital customer experience. *J Consum Res* 53(1):136–161. <https://doi.org/10.1093/jcr/ucaf037>
18. Djafarova E, Trofimenko O (2021) It's so authentic!': the role of perceived authenticity in UGC-driven influencer marketing on TikTok. *J Interact Mark* 56:81–96. <https://doi.org/10.1016/j.intmar.2021.05.002>
19. Dogaru I, Furnham A, McClelland A (2024) Understanding how the presence of music in advertisements influences consumer behaviour. *Acta Psychol* 248:104333. <https://doi.org/10.1016/j.actpsy.2024.104333>
20. Dong X, Liu H, Xi N, Liao J, Yang Z (2024) Short video marketing: what, when and how short-branded videos facilitate consumer engagement. *Internet Res* 34(3):1104–1128. <https://doi.org/10.1108/INTR-02-2022-0121>
21. Farace S, Ordenes FV, Herhausen D, Grewal D, Ruyter KD (2026) Standing out while fitting in: visual design of text overlays in social media communication. *J Mark* 90(1):132–151. <https://doi.org/10.1177/00222429251322773>
22. Ferreira DCS, Oliveira-Castro JM (2011) Effects of background music on consumer behaviour: behavioural account of the consumer setting. *Serv Ind J* 31(15):2571–2585. <https://doi.org/10.1080/02642069.2011.531125>
23. Flanagan AJ, Metzger MJ (2008) The credibility of volunteered geographic information. *GeoJournal* 72(3–4):137–148. <https://doi.org/10.1007/s10708-008-9188-y>
24. Gan J, Shi S, Filieri R, Leung WKS (2023) Short video marketing and travel intentions: the interplay between visual perspective, visual content, and narration appeal. *Tour Manag* 99:104795. <https://doi.org/10.1016/j.tourman.2023.104795>
25. Gibreel O, AlOtaibi DA, Altmann J (2018) Social commerce development in emerging markets. *Electron Commer Res Appl* 27:152–162. <https://doi.org/10.1016/j.elerap.2017.12.008>
26. Gong Y, Liu X (2001) Video summarization with minimal visual content redundancies. In: *Proceedings of the 2001 International Conference on Image Processing*, vol 2. IEEE, pp 362–365. <https://doi.org/10.1109/ICIP.2001.958126>
27. Hautz J, Füller J, Hutter K, Thürndl C (2014) Let users generate your video ads? The impact of video source and quality on consumers' perceptions and intended behaviors. *J Interact Mark* 28(1):1–15. <https://doi.org/10.1016/j.intmar.2013.06.003>
28. Huang L, Dong X, Yuan H, Wang L (2023) Enabling and inhibiting factors of the continuous use of mobile short video APP: satisfaction and fatigue as mediating variables respectively. *Psychol Res Behav Manage* 16:3001–3017.

<https://doi.org/10.2147/PRBM.S411337>

29. Huo H, Zhao J, Zhou J (2026) The influence of virtual anchor language styles on consumer purchase intentions: the role of psychological vigilance and cultural endorsements. *J Retailing Consumer Serv* 91:104750. <https://doi.org/10.1016/j.jretconser.2026.104750>
30. Knight P, Freeman I, Stuart S, Griggs G, O'Reilly N (2014) Semiotic representations of Olympic mascots revisited: virtual mascots of the games 2006–2012. *Int J Event Festival Manage* 5(1):74–92. <https://doi.org/10.1108/IJEFM-03-2012-0010>
31. Kolotouchkina O (2018) Engaging citizens in sports mega-events: the participatory strategic approach of Tokyo 2020 Olympic. *Communication Soc* 31(4):45–58. <https://doi.org/10.15581/003.31.4.45-58>
32. Lee S, Kim E (2024) Attention dynamics in short-form video: a comparative study of professional vs. amateur content. *New Media Soc* 26(3):789–810. <https://doi.org/10.1177/14614448231201543>
33. Li B, Zhao J (2023) Visual-audio correspondence and its effect on video tipping: evidence from Bilibili vlogs. *Inf Process Manag* 60(3):103347. <https://doi.org/10.1016/j.ipm.2023.103347>
34. Li H, Xu J, Fang M, Tang L, Pan Y (2023) A study and analysis of the relationship between visual-auditory logos and consumer behavior. *Behav Sci* 13(7):613. <https://doi.org/10.3390/bs13070613>
35. Li Y, Xie Y (2020) Is a picture worth a thousand words? An empirical study of image content and social media engagement. *J Mark Res* 57(1):1–19. <https://doi.org/10.1177/0022243719881113>
36. Liang A, Sun L, Wang H (2025) A study of the effect of mascot attractiveness on attitudes toward sporting events. *Sci Rep* 15:20385. <https://doi.org/10.1038/s41598-025-09406-z>
37. Ma X, Ren J, Lang X, Yang Z, Li T (2024) The influence of live video hosts' speech act on purchase behaviour. *J Retailing Consumer Serv* 81:103984. <https://doi.org/10.1016/j.jretconser.2024.103984>
38. Ma Z, Gu B (2022) The influence of firm-generated video on user-generated video: evidence from China. *Int J Eng Bus Manage* 14:18479790221118628. <https://doi.org/10.1177/18479790221118628>
39. Malanski D, Peña EF (2022) A Mulatto Brazil: the narrative of the Dionysian character of Brazilian society in the 2014 FIFA World Cup. *Natl Identities* 25(4):423–439. <https://doi.org/10.1080/14608944.2022.2125941>
40. Melzner J, Raghubir P (2023) The sound of music: the effect of timbral sound quality in audio logos on brand personality perception. *J Mark Res* 60(5):932–949. <https://doi.org/10.1177/00222437221135188>
41. Meng L, Kou S, Duan S, Bie Y (2024) The impact of content characteristics of short-form video ads on consumer purchase behavior: evidence from TikTok. *J Bus Res* 183:114874. <https://doi.org/10.1016/j.jbusres.2024.114874>
42. Palladino A, Lee M, Zhou X (2022) An analysis of Olympic mascot design toward attitude and purchase intention. *Int J Sport Communication* 15(3):190–196. <https://doi.org/10.1123/ijsc.2022-0052>
43. Quan Z, Qian W, Mao J (2024) The impact of mascot attributes on sponsorship effectiveness: the case of 'Bing Dwen Dwen' at the Beijing Winter Olympics. *Asia Pac J Mark Logistics* 36(4):1024–1042. <https://doi.org/10.1108/APJML-05-2023-0485>
44. Schivinski B, Dabrowski D (2016) The effect of social media communication on consumer perceptions of brands. *J Mark Commun* 22(2):189–214. <https://doi.org/10.1080/13527266.2013.871323>
45. Scott SP, Sheinin D, Labrecque LI (2022) Small sounds, big impact: sonic logos and their effect on consumer attitudes, emotions, brands and advertising placement. *J Prod Brand Manage* 31(7):1091–1103. <https://doi.org/10.1108/JPBM-06-2021-3507>
46. Shi R, Wang M, Liu C, Gull N (2023) The influence of short video platform characteristics on users' willingness to share marketing information: based on the SOR model. *Sustainability* 15(3):2448. <https://doi.org/10.3390/su15032448>
47. Shih HC (2011) Key-frame extraction and key-frame rate determination using human attention modeling. In: *Proceedings of the 2011 IEEE International Conference on Multimedia and Expo*. IEEE, pp 1–4. <https://doi.org/10.1109/ICME.2011.6012109>

48. Smith AN, Fischer E, Yongjian C (2012) How does brand-related user-generated content differ across YouTube, Facebook, and Twitter? *J Interact Mark* 26(2):102–113. <https://doi.org/10.1016/j.intmar.2012.01.002>
49. Su H, Wang H, Chen Y (2025) The power of cute: how mascot cuteness shapes perception in charity and commercial sporting events. *Sport Mark Q* 34(4):343–356. <https://doi.org/10.1177/10616934251391646>
50. Sunaga T (2018) How the sound frequency of background music influences consumers' perceptions and decision making. *Psychol Mark* 35(4):253–267. <https://doi.org/10.1002/mar.21084>
51. Sweller J (1988) Cognitive load during problem solving: effects on learning. *Cogn Sci* 12(2):257–285. https://doi.org/10.1207/s15516709cog1202_4
52. Tao Y, Ren Z, Liu X, Chen Y (2026) Color, voice type and pitch: how multi-sensory cues of virtual shopping assistants influence young people's purchase intention in VR shopping. *Electron Commer Res* 26(2):1853–1885. <https://doi.org/10.1007/s10660-025-09965-3>
53. Tavanapong W, Zhou J (2004) Shot clustering techniques for story browsing. *IEEE Trans Multimedia* 6(4):517–527. <https://doi.org/10.1109/TMM.2004.830810>
54. Tennent KD, Gillett AG (2018) Opportunities for all the team: entrepreneurship and the 1966 and 1994 Soccer World Cups. *Int J Hist Sport* 35(7–8):767–788. <https://doi.org/10.1080/09523367.2018.1544555>
55. Truong BT, Venkatesh S (2007) Video abstraction: a systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications* 3(1):3. <https://doi.org/10.1145/1198302.1198305>
56. Vazquez EE, Patel C, Alvidrez S, Siliceo L (2023) Images, reviews, and purchase intention on social commerce: the role of mental imagery vividness, cognitive and affective social presence. *J Retailing Consumer Serv* 74:103415. <https://doi.org/10.1016/j.jretconser.2023.103415>
57. Wang ZC, Jiang QL (2024) A study on the willingness of 'Generation Z' consumers to use online virtual try-on shopping services based on the S-O-R framework. *Systems* 12(6):217. <https://doi.org/10.3390/systems12060217>
58. We Are Social and Meltwater (2025) Digital 2025: global overview report. DataReportal. <https://datareportal.com/reports/digital-2025-global-overview-report>. Accessed 6 Sept 2026
59. Wu X, Yang C (2025) From symbols to strategy: the role of mascot design in sports management and cultural communication. *OALib* 12(11):1–7. <https://doi.org/10.4236/oalib.1114242>
60. Xiao Q, Huang W, Qu L, Li X (2025) The impact of multimodal information features of short sales videos on consumer engagement behavior: a multi-method approach. *J Retailing Consumer Serv* 82:104136. <https://doi.org/10.1016/j.jretconser.2024.104136>
61. Xiong Y, Wei N, Qiao K, Li Z, Li Z (2024) Exploring consumption intent in live e-commerce barrage: a text feature-based approach using BERT-BiLSTM model. *IEEE Access* 12:69288–69298. <https://doi.org/10.1109/ACCESS.2024.3399095>
62. Xu W, Zhang X, Chen R, Yang Z (2023) How do you say it matters? A multimodal analytics framework for product return prediction in live streaming e-commerce. *Decis Support Syst* 172:113984. <https://doi.org/10.1016/j.dss.2023.113984>
63. Yang Q, Han Q, Zhang T, Lin Y (2026) Process reveal or product display? The impact of product presentation strategy on consumer engagement in short-video marketing. *J Retailing Consumer Serv* 92:104827. <https://doi.org/10.1016/j.jretconser.2026.104827>

Figures

北京2022年冬奥会和冬残奥会吉祥物宣传片

124.4万 4137 2019-09-18 11:10:11

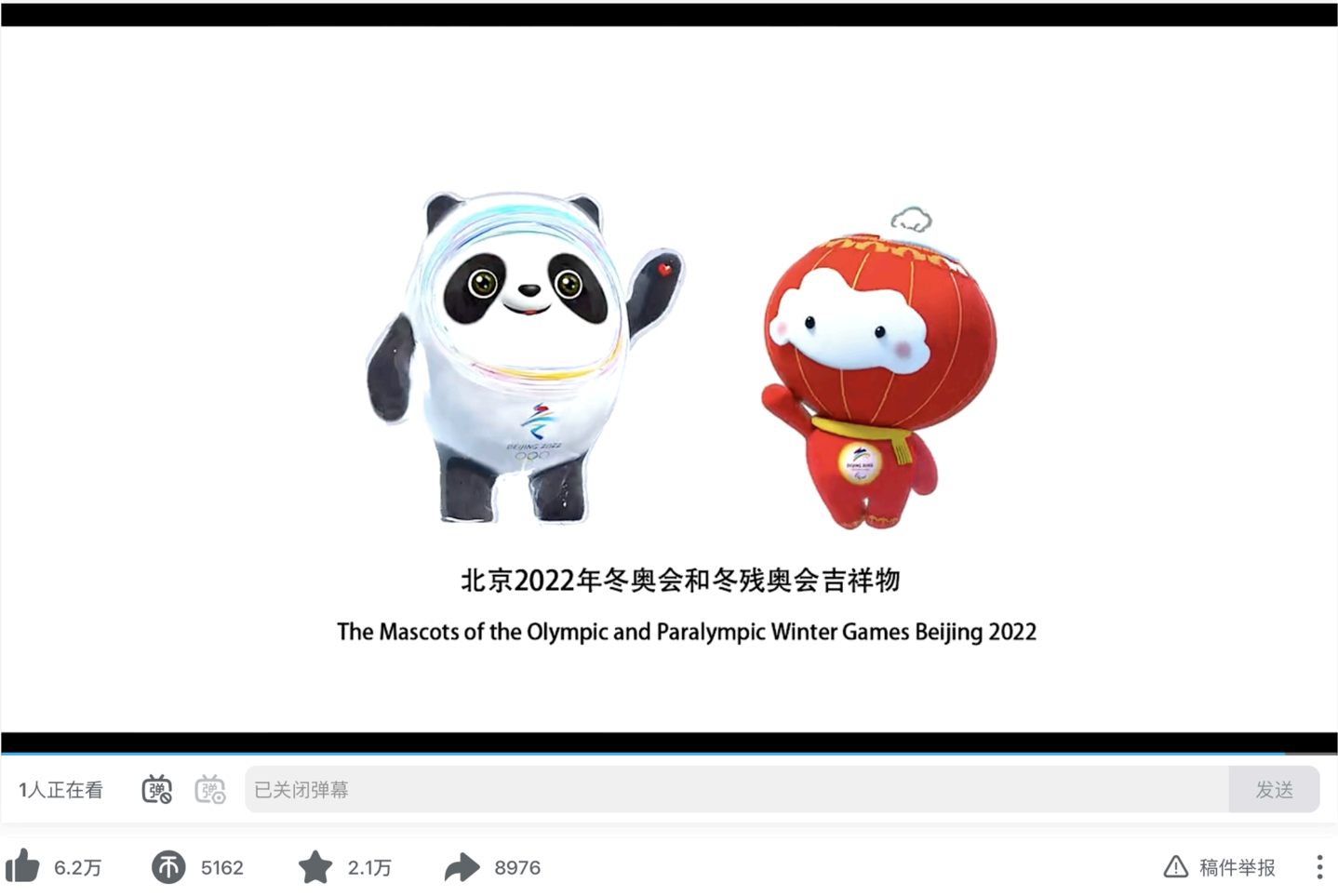


Figure 1

Representative sports event mascot short-form video sample

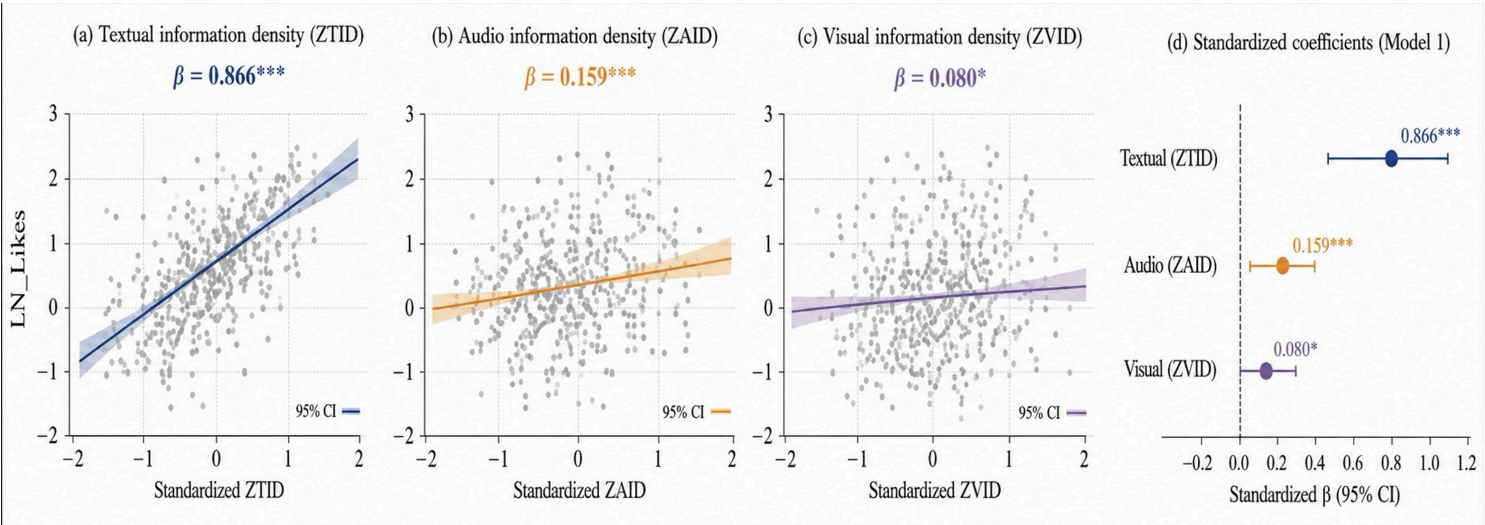


Figure 2

Main effects of multimodal information density on liking behavior

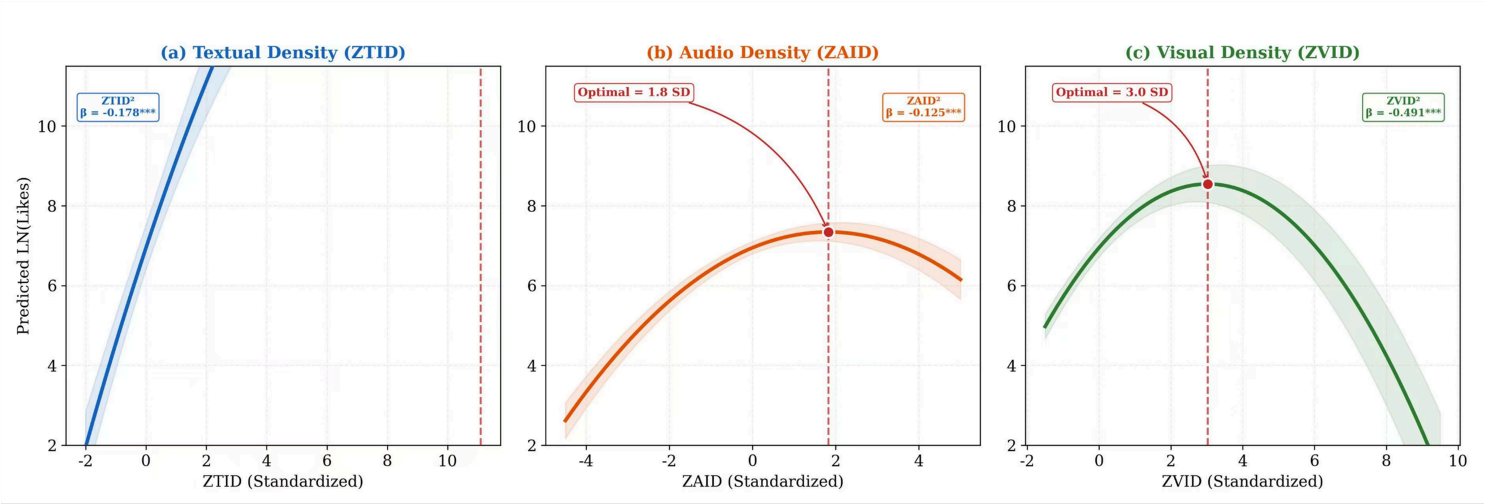


Figure 3

Inverted U-Shaped Marginal Effects of Multimodal Information Density on Liking Behavior

Inverted U-shaped marginal effects of multimodal information density on liking behavior. Sub-figures (a), (b), and (c) correspond respectively to the predicted like-value curves for textual, audio, and visual information density. The red dashed line marks the optimal density threshold for each modality.

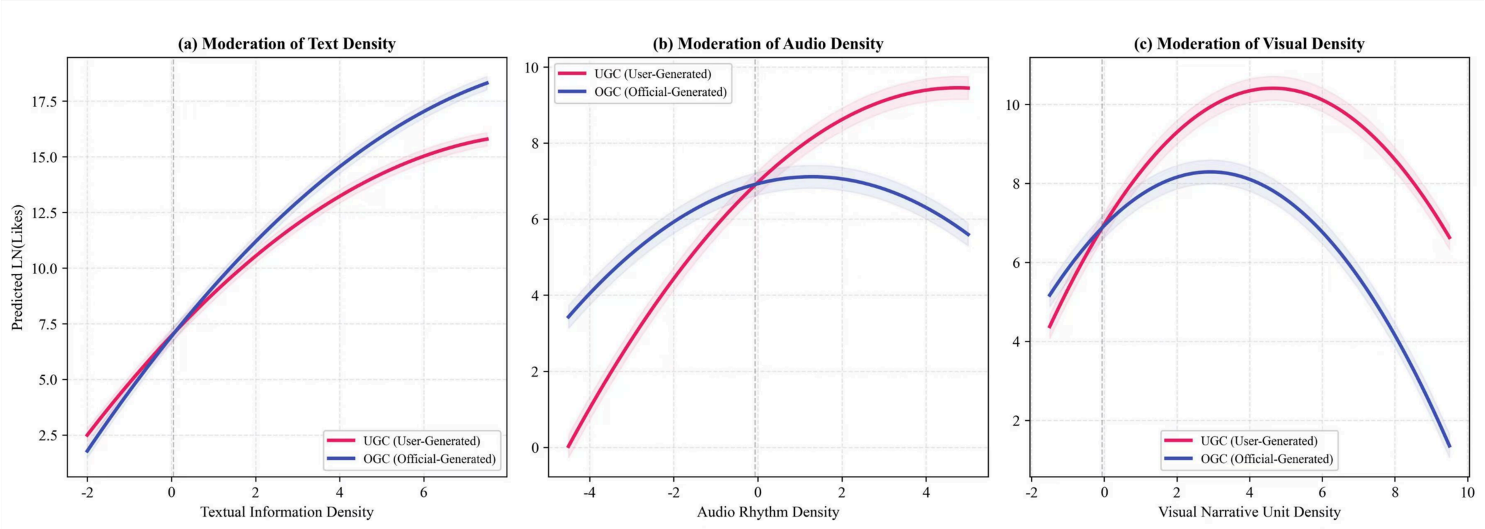


Figure 4

Moderation Effects of Content Source Type (OGC vs UGC)

Moderating effects of video source type (OGC vs. UGC). The red line represents UGC (user-generated content), and the blue line represents OGC (Officially generated content). Textual density has a stronger positive effect in OGC contexts, whereas audio and visual density exhibit stronger associations with liking behavior in UGC contexts, confirming the differential moderating role of content source.

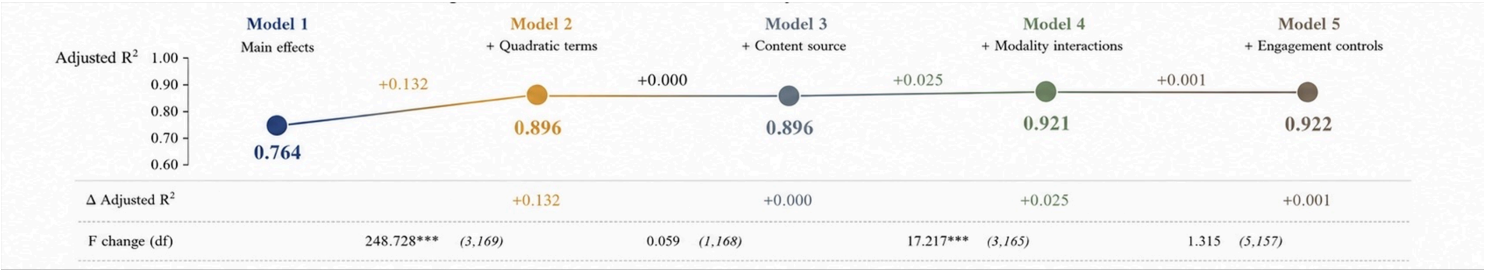


Figure 5

Hierarchical increase in adjusted R^2 across nested models

Change in adjusted R^2 across hierarchical regression nested models. Main effects, quadratic nonlinear terms, content source, modality interaction terms, and additional interaction control variables are sequentially incorporated. Model fit improves significantly after adding quadratic terms and interaction terms, whereas the incremental explanatory power contributed by additional interaction control variables alone is marginal and non-significant, confirming the robustness of the core conclusions.