Journal of
Theoretical and Applied Electronic Commerce Research

MDPI

Article

Striking the Right Pitch: The Inverted U-Shaped Effect of AI Anchor Pitch Variability on Consumer Engagement

Xiaochen Liu [1], Qiang Yang [2] and Yushi Jiang [3,*]

[1] School of Economics and Management, Liaoning University of Technology, Jinzhou 121001, China; claireliu2023@gmail.com
[2] School of Business, Nanjing Audit University, Nanjing 211815, China; yangqiang@nau.edu.cn
[3] School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, China
* Correspondence: 771844204@my.swjtu.edu.cn

Abstract

As artificial intelligence and digital human technologies become increasingly integrated into livestream commerce, AI anchors are becoming important marketing agents. Yet prior research has focused primarily on their visual characteristics, leaving dynamic vocal cues largely unexplored. Drawing on social response theory and perceived authenticity research, this study examines the nonlinear association between AI anchor pitch variability and consumer engagement, together with a proposed psychological pathway and boundary condition. Study 1 analyzes 4322 product-presentation segments nested within 330 AI-anchored livestreams and 85 independent accounts on Douyin. Negative binomial models, formal boundary-slope tests, and additional specifications using account and livestream-session fixed effects, a correlated-random-effects decomposition, and viewer-minutes exposure provide robust evidence of an inverted U-shaped association between pitch variability and real-time danmaku engagement. Evidence concerning appearance-realism moderation is conditional and specification-sensitive across alternative pitch operationalizations, exposure definitions, and within-account specifications. Study 2 uses a preregistered multi-stimulus mixed design with four AI anchors, four products, and three between-participants pitch-variability conditions. Correctly scaled planned contrasts show that moderate pitch variability produced greater perceived authenticity and engagement intentions than the average of the two endpoint conditions. A 2-1-1 multilevel analysis yielded an indirect-effect pattern consistent with the proposed role of perceived authenticity. Models allowing treatment effects to vary across the 16 included anchor-product combinations showed a positive average moderate-pitch advantage, although its magnitude varied across stimuli. These findings extend livestream-commerce research from human streamers to AI-mediated communication while indicating that appearance-realism moderation, stimulus-level generalization, and causal mediation require further replication.

Keywords: AI anchor; pitch variability; consumer engagement; perceived authenticity; appearance realism; livestream commerce

Check for updates

Academic Editors: Chenglu Wang, Zhen Li, Xiaoling Li, Hongfei Liu and Morgan Yang

Received: 22 June 2026
Revised: 9 August 2026
Accepted: 11 August 2026
Published: 14 August 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid development of generative artificial intelligence, speech synthesis, and digital human technologies is bringing AI anchors into the routine operations of livestream commerce [1,2]. Compared with human streamers, AI anchors offer several operational advantages, including stable online availability, lower operating costs, and standardized

service processes [3]. They are therefore increasingly used for product presentations, interactive responses, and long-duration livestreaming tasks [4–6]. However, an AI anchor is not merely an information display tool. It appears in livestream rooms with human-like appearance, voice, and language, and viewers rely on these cues to judge whether it seems natural, authentic, trustworthy, and interactive [7,8]. As these systems increasingly mediate commercial communication at scale, their expressive design may shape not only firm performance but also consumers' trust, attention, and willingness to participate in AI-mediated marketplaces [9,10]. Understanding when synthesized vocal expression facilitates rather than undermines engagement is therefore both a marketing issue and a broader consumer-welfare concern.

Among the various expressive cues of AI anchors, voice plays a particularly important role. Livestream commerce relies heavily on continuous product explanation and real-time interaction. Viewers not only observe the product and the anchor's visual presentation but also continuously receive vocal information from the anchor [11,12]. Vocal features such as pitch, speech rate, pauses, loudness, and pitch variation can shape viewers' judgments of the anchor's expressive state, communicative intent, and interaction atmosphere [12,13]. This is especially important for AI anchors. Their voice not only conveys product information but also directly affects whether viewers perceive them as natural and human-like. If an AI anchor's voice is too flat, viewers may perceive it as mechanical narration. If its voice changes too dramatically, viewers may find it unnatural, exaggerated, or overly imitative of human expression. Thus, AI anchor voice design has become an important marketing issue that may influence consumer engagement.

However, previous research on AI marketing and virtual streamers has focused more on visual cues, such as visual attractiveness, facial expressions, and identity disclosure [14]. Prior studies have shown that the appearance of virtual humans or AI service agents can influence consumers' anthropomorphic perceptions, trust, social presence, and behavioral responses [15]. In contrast, the vocal features of AI anchors, especially the dynamic changes in voice over time, remain insufficiently examined. This gap is particularly important in livestream commerce because livestream interaction is not a static viewing experience. Rather, it is a dynamic communication process shaped by continuous vocal explanation, product presentation, and viewer feedback. Understanding AI anchors only through the lens of visual realism is therefore insufficient to explain how viewers form authenticity judgments and consumer engagement.

To address this gap, this study focuses on AI anchor pitch variability. Pitch variability refers to the degree to which an anchor's vocal pitch fluctuates during continuous speech [11]. For human streamers, pitch variability is often associated with emotional expression, vocal energy, and interaction enthusiasm [13]. In the context of AI anchors, however, pitch variability has a different meaning. It does not directly reflect genuine emotion. Rather, it is a vocal cue generated by speech synthesis systems, scripted content, and algorithmic parameters [16]. Therefore, the effect of pitch variability may be more complex for AI anchors. On the one hand, moderate pitch variability can reduce the sense of mechanical narration and make an AI anchor sound more natural and expressive. On the other hand, excessive or poorly coordinated pitch variability may reveal the artificial nature of the voice and make viewers perceive it as inauthentic or unnatural. This raises a central question: does greater pitch variability always increase consumer engagement, or does consumer engagement peak at an intermediate level within the range examined?

To answer this question, this study draws on social response theory and perceived authenticity research to examine the effect of AI anchor pitch variability on consumer engagement. We argue that although viewers know AI anchors are not human, they may still respond to them socially based on their vocal and visual cues. Moderate pitch vari-

ability may enhance viewers' perceptions of the AI anchor's naturalness and authenticity, thereby promoting consumer engagement. At the same time, the visual appearance of an AI anchor may shape how viewers interpret its voice. In particular, when an AI anchor has a high level of appearance realism, viewers may hold stronger expectations regarding the naturalness of its voice. Accordingly, this study further examines appearance realism as a boundary condition for the effect of AI anchor voice.

This study tests the proposed framework through two complementary studies. Study 1 uses real-world product presentation segments nested within livestreams and accounts to examine whether pitch variability has an inverted U-shaped association with observed danmaku engagement and whether appearance realism moderates this association. Study 2 uses a multi-stimulus mixed design that manipulates pitch variability between participants while varying AI anchor identity and product within participants; it examines perceived authenticity and self-reported consumer engagement intention. The two studies therefore provide complementary rather than directly replicative evidence: Study 1 offers ecological validity for actual engagement behavior, whereas Study 2 provides stronger control over the vocal manipulation but examines self-reported behavioral intentions.

This study makes several theoretical contributions. First, it extends AI anchor research from visual realism to dynamic vocal cues, responding to the multimodal nature of AI-enabled livestream marketing. Second, by focusing on pitch variability, this study shows that AI anchor voice design is not simply a matter of making the voice more human-like, more exaggerated, or more variable. Instead, the findings suggest that an intermediate level of pitch variability may generate more favorable responses within the observed and experimentally tested ranges. Third, this study explains the psychological process through which AI anchor voice affects consumer engagement from the perspective of perceived authenticity. By incorporating appearance realism into the framework, it also offers a new perspective on the joint role of visual and vocal cues in AI-mediated livestream interaction. In practical terms, the findings provide useful implications for brands, platform operators, and AI anchor service providers.

2. Literature Review

2.1. Research on AI Anchors

AI anchors refer to digital anchors powered by artificial intelligence, virtual avatars, speech synthesis systems, and automated scripts that perform product presentations, promotional announcements, interactive responses, and shopping guidance in livestreaming settings [17,18]. Unlike traditional virtual images or static digital humans, AI anchors are capable of continuous output, near-real-time interaction, and standardized communication. They can provide sustained explanations and responses based on product information, promotional rhythms, and viewer feedback [19,20]. Thus, an AI anchor is not only the front-end representation of an intelligent technological system but also a marketing agent that consumers directly encounter, perceive, and evaluate in livestream commerce [6,21].

Previous research on AI anchors can be broadly divided into three streams. The first stream examines consumer acceptance and adoption of AI anchors. These studies usually draw on technology acceptance, perceived value, social presence, and trust to explain whether consumers are willing to watch AI-anchored livestreams [22,23]. For example, Liu et al. (2025) examined AI-generated virtual anchor interaction and found that AI anchors can influence purchase intention through social presence and perceived value [17]. Shui et al. (2025), drawing on the stimulus–organism–response framework, showed that AI anchor characteristics, product type, and livestreaming context affect purchase intention through consumer trust [18]. The second stream compares AI anchors with human streamers. Related studies suggest that human streamers are more likely to build emotional

trust through emotional expression and personal experience, whereas AI anchors have advantages in automation and efficiency but may be limited in emotional expression, which can weaken consumers' emotional connection and trust [17,24]. Similarly, Yoo et al. (2025) compared AI influencers with human influencers and found that AI influencers can, in some contexts, generate effects comparable to those of human influencers, although their persuasiveness still depends on psychological distance, authenticity judgments, and perceived human likeness [25]. The third stream focuses on the characteristics, image, and interaction mechanisms of AI anchors. For example, Li et al. (2024) found that virtual streamer anthropomorphism influences consumer purchase intention through cognitive trust, suggesting that consumers form trust judgments based on whether a virtual streamer has human-like appearance, expression, and interaction capabilities [26]. Research on the sociality of virtual livestream anchors also shows that highly social virtual anchors generate greater consumer experience value than less social ones [14,27,28].

Overall, previous research has mainly examined why consumers accept AI anchors and whether AI anchors can enhance trust, purchase intention, and sales performance. However, less attention has been paid to how specific expressive cues of AI anchors affect consumers' real-time engagement. In livestream commerce, consumer engagement is reflected not only in purchase intention or purchase behavior but also in immediate interactive behaviors such as likes, danmaku comments, continued viewing, and follows [19,20]. In addition, prior research has placed greater emphasis on AI anchors' visual image and anthropomorphic features, while paying relatively limited attention to their vocal features [29]. In fact, consumers' judgments of whether an AI anchor seems natural, authentic, and interactive are shaped to a large extent by the anchor's continuous vocal presentation. Based on these gaps, this study shifts the focus from visual realism and general acceptance of AI anchors to their dynamic vocal cues, examining how AI anchor pitch variability influences consumer engagement. In doing so, this study enriches research on the multimodal cues and marketing effects of AI anchors.

2.2. Language and Vocal Cues in Livestream Commerce

Livestream commerce is a marketing context that relies heavily on verbal communication and real-time interaction [30]. Streamers introduce product information through continuous explanation, respond to viewer questions through interactive language, create an engaging atmosphere through emotional expression, and stimulate consumer decisions through promotional messages [31].

Previous research on expressive cues in livestream commerce can be divided into two broad streams. The first stream focuses on language cues, including streamers' message content, interactive language, and persuasive strategies. For example, Chen et al. (2023), drawing on speech act theory, argued that streamers do more than transmit product information; they also influence consumer decisions through different speech acts such as commissives, directives, expressives, and representatives [30]. Ma et al. (2024) further found that commissive speech acts positively influence purchase behavior, whereas representative and expressive speech acts may have negative effects; customer engagement also mediates some of the relationships between speech acts and purchase behavior [32]. Luo et al. (2021) examined persuasive language styles in livestream commerce and found that different linguistic persuasion styles affect product sales [33]. More recent research has also examined combinations of language strategies, showing that reward-based, exaggerated, and credibility-related persuasive language strategies can shape customer engagement and product sales [34].

The second stream focuses on vocal cues in livestream commerce, including speech rate, pitch, loudness, pauses, tone, and prosody. Compared with language content, vocal

cues function as paralinguistic or nonverbal cues. They do not directly change the information conveyed by the streamer, but they shape viewers' judgments of the streamer's emotional state, professionalism, warmth, and credibility [11,13]. In recent years, researchers have begun to examine the marketing effects of streamer vocal features using real livestream data and acoustic analysis. Luo et al. (2025), using Douyin livestream data, extracted streamers' pitch, loudness, and speech rate through acoustic analysis and found that pitch and speech rate positively influence sales performance, whereas loudness has a negative effect [12]. They also found that emotional vocal cues such as pitch and loudness are more effective for experience products, whereas informational vocal cues such as speech rate are more effective for search products. Li et al. (2026), in a tourism livestreaming context, examined vocal cues and in-process engagement and found an inverted U-shaped effect of speech rate on consumer engagement [24].

However, research on language and vocal cues in livestream commerce still has two limitations. First, prior studies have mainly focused on human streamers, with limited attention to the language and vocal cues of AI anchors. Human streamers and AI anchors differ in the source of their expression, the way consumers interpret their communication, and the psychological mechanisms through which their cues operate. Therefore, findings from human streamer research cannot be directly transferred to the AI anchor context. Second, previous research on vocal cues has focused more on relatively static or overall acoustic indicators, such as speech rate, mean pitch, and loudness, while paying less attention to how voice changes over time. In livestream commerce, a streamer's voice is not an isolated segment but unfolds continuously during product explanation and real-time interaction. Based on these considerations, this study focuses on AI anchor pitch variability as a dynamic vocal cue and examines how it influences consumer engagement. This perspective provides a new way to understand multimodal interaction and consumer engagement in AI-mediated livestream commerce.

3. Research Hypotheses

3.1. AI Anchor Pitch Variability and Consumer Engagement

Social response theory suggests that individuals respond socially to technologies when those technologies present human-like social cues, even when individuals clearly know that the technologies are not human [35,36]. In AI-anchored livestreaming, voice is a highly salient social cue that consumers continuously receive. Research on vocal cues in livestream commerce shows that acoustic features such as pitch, loudness, and speech rate can influence viewer responses and sales performance [12]. Research on voice naturalness also suggests that the perceived naturalness of a voice shapes how people interact with artificial agents and synthetic voices [37,38]. In addition, research on the vocal uncanny valley shows that when a voice deviates from familiar natural voice patterns, it may trigger perceptions of unnaturalness and eeriness [39,40].

Pitch variability refers to the degree to which an AI anchor's vocal pitch fluctuates during continuous speech. Low pitch variability usually indicates limited vocal change, weak rhythm, and insufficient expressive layering [13]. In livestream commerce, where continuous explanation and real-time interaction are central, an overly flat voice may make an AI anchor seem mechanical, monotonous, and lacking in interactive warmth [38]. In this case, consumers may perceive the AI anchor as a background system reading from a script rather than as an interaction partner worth responding to, thereby reducing their willingness to like the livestream or send danmaku comments.

As pitch variability increases, the AI anchor's voice becomes more rhythmically varied and expressive. Moderate pitch variability can make product explanations more vivid, highlight promotional messages more clearly, and make interactive responses closer to

natural everyday communication. For consumers, such moderate vocal variation can increase the attractiveness and vividness of the voice, making them more willing to express immediate approval through likes and to ask questions, comment, or express emotions through danmaku comments.

However, pitch variability is not always beneficial. For AI anchors, excessive pitch variability may no longer be interpreted as natural expression but instead as exaggerated, deliberate, or poorly coordinated algorithmic performance. This problem may be especially salient when the vocal fluctuations do not match the semantic emphasis, product attributes, or atmosphere of the livestream. Excessive pitch variability may reveal the synthetic nature of the AI voice and create a sense of mismatch or distance. As a result, consumer engagement may decline.

Therefore, this study proposes that AI anchor pitch variability has an inverted U-shaped relationship with consumer engagement. Low pitch variability may fail to stimulate interaction, moderate pitch variability may promote consumer engagement, and excessive pitch variability may reduce consumer engagement because it appears unnatural or overly performative. Thus:

H1. *AI anchor pitch variability has an inverted U-shaped relationship with consumer engagement. As pitch variability increases from a low to a moderate level, consumer engagement increases; when pitch variability exceeds a moderate level, consumer engagement decreases.*

3.2. AI Anchor Pitch Variability and Perceived Authenticity

Perceived authenticity is an important consumer judgment of whether a marketing agent appears natural, trustworthy, and non-manipulative. Authenticity research suggests that consumers may judge authenticity based on either an object's similarity to a prototype or the extent to which its expression appears natural, sincere, and credible [28,41]. In virtual influencer research, authenticity is also a key factor explaining consumer trust, acceptance, and interaction responses [27,42]. For AI anchors, consumers usually know that the anchor is not a real person. Therefore, authenticity does not mean being "truly human"; rather, it refers to whether the AI anchor's expression seems natural, coherent, credible, and not rigid.

Pitch variability can shape consumers' perceived authenticity of an AI anchor. Low pitch variability can make the voice sound programmatic and template-like, leading consumers to feel that the AI anchor is merely executing a script [37]. Moderate pitch variability, in contrast, adds rhythm and expressive layering to the voice, making the AI anchor's explanation closer to natural communication and thereby strengthening consumers' authenticity judgments. However, when pitch variability is too high, the voice may appear overly emotional or excessively human-like, which may draw attention to its artificial origin. Research on voice naturalness and the vocal uncanny valley suggests that the perceived naturalness of synthetic voices depends not only on whether the voice sounds human-like but also on whether its prosody, pitch variation, and expressive pattern fit listeners' expectations of natural communication [37–40]. Thus, excessive or poorly coordinated pitch variability may weaken perceived authenticity. Therefore,

H2. *AI anchor pitch variability has an inverted U-shaped relationship with perceived authenticity. Specifically, moderate pitch variability leads to higher perceived authenticity than either low or high pitch variability.*

3.3. The Proposed Indirect Role of Perceived Authenticity

Perceived authenticity not only shapes consumers' evaluations of a marketing agent but also influences their behavioral responses. Influencer marketing research shows that

when consumers perceive an influencer's expression as more authentic and natural, they are more likely to develop trust and respond positively [27,28]. This study argues that AI anchor pitch variability influences consumer engagement because it changes consumers' judgments of the authenticity of the AI anchor's expression. Moderate pitch variability makes the AI anchor's voice closer to natural communication and strengthens consumers' perceptions of naturalness and credibility.

By contrast, when pitch variability is too low, the AI anchor may be perceived as mechanically reading a script; when pitch variability is too high, it may be perceived as exaggerated or fake algorithmic performance. In both cases, consumers' authenticity judgments decline, which in turn reduces their willingness to engage. Thus, pitch variability affects consumer engagement not merely through the acoustic stimulus itself but by shaping whether consumers perceive the AI anchor as natural, authentic, and credible. Therefore,

H3. *Moderate pitch variability, relative to low or high variability, is associated with higher consumer engagement through higher perceived authenticity; accordingly, a significant indirect effect through perceived authenticity is expected.*

3.4. The Moderating Role of Appearance Realism

Livestream commerce is a typical multimodal interaction context. Consumers not only hear an AI anchor's voice but also see its appearance, facial expressions, lip movements, and body movements [43]. Research on virtual agents and service robots suggests that human likeness influences consumers' social judgments, trust, and behavioral responses toward artificial agents [44,45]. At the same time, research on the uncanny valley and multimodal consistency shows that inconsistency among appearance, voice, and behavioral cues may lead to feelings of mismatch, eeriness, and reduced trust [39,43,46]. Therefore, the effect of AI anchor vocal cues does not occur in isolation but is shaped by visual realism.

When appearance realism is high, moderate pitch variability is more likely to be consistent with the AI anchor's human-like appearance. In this case, vocal variation can complement visual realism and make the AI anchor appear more natural, expressive, and interactive. This visual–vocal consistency can strengthen consumers' positive responses to the AI anchor. Therefore, under high appearance realism, moderate pitch variability may have a stronger positive effect on consumer engagement, including likes and danmaku comments.

However, high appearance realism may also raise consumers' expectations regarding voice naturalness. When an AI anchor looks highly human-like, consumers expect its voice to sound similarly natural. If pitch variability is excessive, exaggerated, or inconsistent with the semantic context, consumers may more easily perceive a mismatch between the AI anchor's appearance and voice, which can create a sense of incongruity. In this sense, high appearance realism may amplify both the positive effect of moderate pitch variability and the negative effect of excessive pitch variability, making the inverted U-shaped relationship between pitch variability and consumer engagement more pronounced.

In contrast, when appearance realism is low, consumers may have lower expectations regarding the naturalness of the AI anchor's voice. Because a low-realism appearance already signals the anchor's nonhuman nature, consumers may be more tolerant of mechanical or imperfect vocal features. In this case, the effect of pitch variability on consumer engagement may be weaker. Therefore,

H4. *Appearance realism moderates the inverted U-shaped relationship between AI anchor pitch variability and consumer engagement. Compared with low appearance realism, high appearance realism strengthens the inverted U-shaped relationship between pitch variability and consumer engagement.*

The conceptual model of this study is presented in Figure 1.

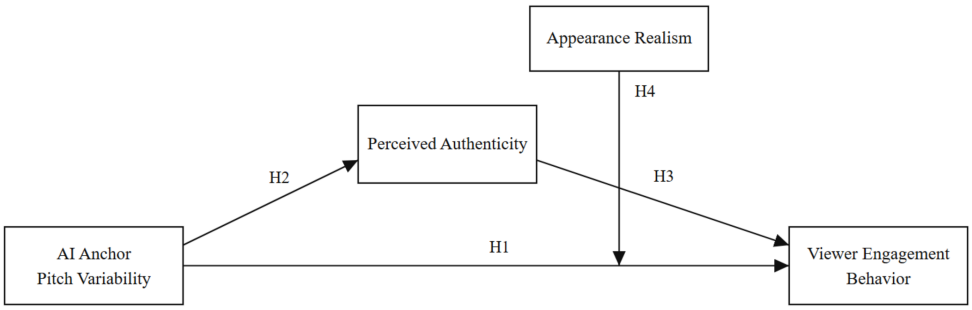

Figure 1. Conceptual Model.

4. Study 1

Study 1 uses real consumer behavior data, livestream videos, and AI anchor speech content from Douyin livestream commerce to test the proposed theoretical model. Douyin is one of China's leading livestream commerce platforms, with an active user base and a relatively mature livestream shopping ecosystem. Compared with an experimental setting, platform data more directly capture consumers' immediate interactive behavior during AI-anchored livestreams and therefore provide an appropriate context for examining the relationship between AI anchors' vocal characteristics and consumers' real-time engagement.

4.1. Data Sources and Processing

Study 1 used segment-level data obtained through Douchacha, a third-party data service covering publicly accessible Douyin livestreams. The livestream observation period was 1 March to 31 May 2026, and data extraction, cleaning, and archiving were completed on 22 July 2026. The final sample comprised 330 commercial livestreams from 85 independent accounts, each corresponding to one AI anchor; no anchor appeared across multiple accounts. The 4322 retained product-presentation segments were distributed across food and beverages (n = 1161), household products (n = 957), beauty and personal care (n = 863), apparel and accessories (n = 504), electronics (n = 326), health and wellness (n = 290), and other products (n = 221). Further details on the sample hierarchy and account characteristics are presented in Table 1.

We segmented each livestream according to the actual presentation period for each product. A new product presentation segment was identified when the product card changed, the AI anchor began introducing a new product, or the speech transcript contained an explicit cue indicating a product switch. To ensure sample representativeness and comparability, we first stratified the data by follower count, account type, and AI anchor presentation format. We then retained only segments that contained a clearly identifiable product presentation, showed the AI anchor continuously on screen, and had synchronized audio and video. Finally, we removed segments without human-like speech, segments with prolonged intervention by a human co-host, and product presentation segments shorter than 10 s. After screening, the final dataset contained 4322 valid product-level livestream video segments nested within 330 livestreams and 85 accounts. Each livestream contributed 10–18 valid segments, and no single account contributed more than 1.48% of the full sample, indicating that the sample was not disproportionately driven by a small number of accounts. Product presentation segments averaged 85.52 s in length, with a median of 70.64 s. Segmenting by the actual product presentation period, rather than by fixed one-minute intervals, more accurately aligns the AI anchor's vocal delivery, the product being displayed, and consumer engagement behavior, thereby reducing measurement error caused by overlapping product content.

Table 1. Sample Hierarchy and Account Characteristics.

Variable	Category/Statistic	Count
Independent accounts	Total	85
AI anchors	One anchor per account	85
AI anchor gender presentation	Masculine appearance	18
	Feminine appearance	67
Follower base	Below 100,000	14
	100,000–1 million	47
	1–10 million	20
	Above 10 million	4
Sales reputation	Average	14
	Good	53
	Very good	18
Product-presentation segments	Food and beverages	1161
	Household products	957
	Beauty and personal care	863
	Apparel and accessories	504
	Electronics	326
	Health and wellness	290
	Other	221

For each product-presentation segment, we retained structured variables required for analysis, including segment duration, product price and category, concurrent viewers, and the number of newly posted danmaku comments. We also retained anchor-, account-, and livestream-level variables, including anonymized anchor characteristics, follower count, sales reputation, livestream frequency, and time of broadcast. Account and anchor identifiers were replaced with study-specific anonymous codes before analysis.

The analytical dataset contained only aggregate danmaku counts; no danmaku text, viewer usernames, viewer identifiers, or other personally identifiable audience information was collected or retained. The source livestreams were publicly accessible, and the data were processed solely for academic research in accordance with applicable platform and institutional data-governance requirements.

AI-anchored speech was automatically transcribed into Chinese using iFlytek Tingjian. The resulting text was standardized by removing nonsemantic punctuation and harmonizing full-width and half-width characters. We then used the Jieba tokenizer to segment the standardized Chinese transcripts and applied the Linguistic Inquiry and Word Count dictionary to derive the proportions of function words, cognitive-process words, social words, and money-related words. The transcripts were also used to calculate speech rate.

For the audio data, we first converted each product presentation segment into a 16 kHz mono WAV file and then used the openSMILE toolkit to extract fundamental acoustic features of the AI anchor's voice. openSMILE is an open-source acoustic feature extraction toolkit commonly used in speech emotion recognition and business voice analytics; it reliably extracts multidimensional indicators such as fundamental frequency, loudness, spectral features, and prosody [13,47]. Following prior acoustic research, we used the eGeMAPS feature set to extract pitch, loudness, speech rate, and related vocal features. Developed by experts in speech and affective computing, eGeMAPS is suitable for studying emotional expression, social signals, and human–computer interaction [48].

For the video data, we extracted frames from each AI-anchored livestream at one frame per second and used computer vision methods to identify the AI anchor's facial region, facial expressions, and relevant technical image features. To reduce measurement error attributable to any single frame, we averaged the recognition results across all valid

frames within each product presentation segment. Figure 2 presents the data processing procedure for Study 1.

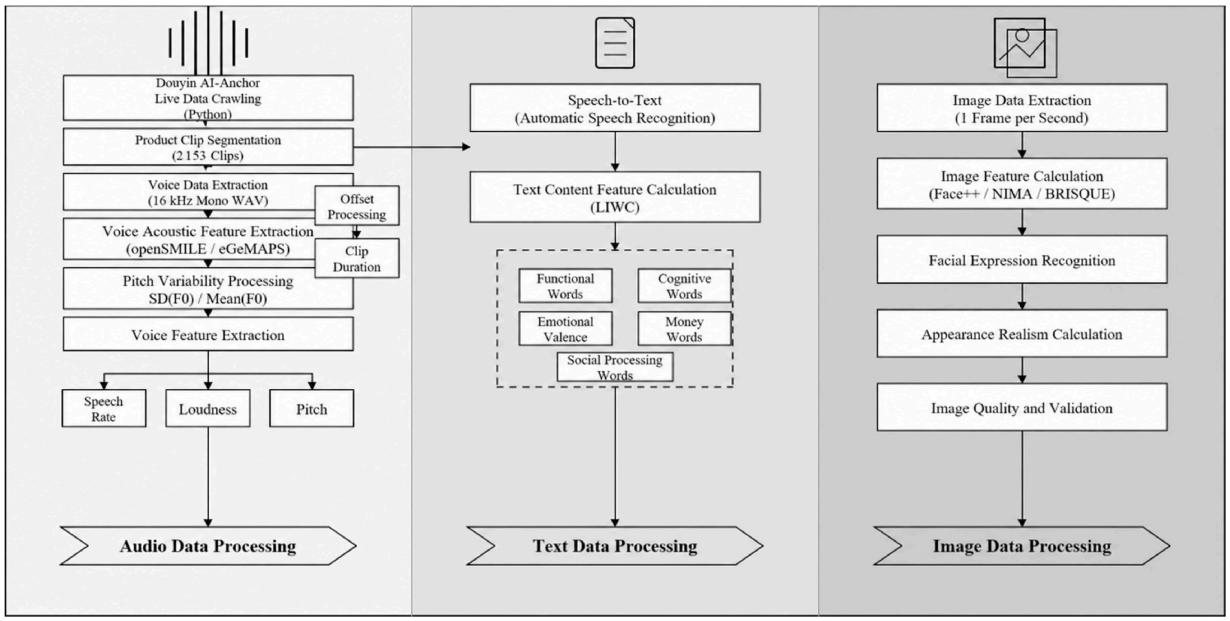

Figure 2. Data Processing Procedure for Study 1.

4.2. Measurement of Variables

4.2.1. Independent Variable

The core independent variable in Study 1 was AI anchor pitch variability, defined as the extent to which the anchor's vocal pitch fluctuated during a product presentation. Each segment was converted to a 16 kHz mono WAV file and analyzed frame by frame using a 25 ms window and a 10 ms frame shift. Fundamental frequency (F0) and voicing probability were extracted with openSMILE's subharmonic-summation algorithm. The prespecified F0 search range was 50–600 Hz, and only frames with a voicing probability of at least 0.70 and F0 within this range were retained. To identify octave and half-octave errors, each candidate F0 was compared with the local median of adjacent valid frames. A frame was corrected when its deviation exceeded 9 semitones, and multiplying or dividing the value by two reduced the deviation to no more than 3 semitones; frames that could not be corrected reliably were set to missing. We additionally removed observations exceeding three robust median-absolute-deviation units from the segment-level semitone median. A segment was eligible only if it contained at least 200 valid F0 frames, and valid frames constituted at least 20% of all analyzed frames. Valid F0 values were converted to a semitone scale, and pitch variability was operationalized as the within-segment standard deviation of semitone F0. All thresholds were specified before model estimation and applied uniformly to the full sample.

4.2.2. Dependent Variable

Consumer engagement was measured as the number of new danmaku comments posted by viewers during each product presentation segment. Danmaku comments are an important indicator of consumers' real-time participation, interactive expression, and social response in livestream commerce and directly reflect viewers' immediate reactions to the current product presentation and the AI anchor's communication style. Because the unit of analysis was the product presentation segment and segment duration varied across products, the natural logarithm of segment duration was included as an offset term to account for differences in exposure time.

4.2.3. Moderator

Appearance realism refers to the extent to which an AI anchor's visual presentation resembles a real human and appears visually natural and human-like. For each of the 85 AI anchors, the research team prepared three silent, complete product-presentation videos, yielding 255 rating materials. These materials were full product-presentation videos rather than short excerpts. For each anchor, three eligible videos were selected without replacement using a computer-generated random sequence. Videos with substantial facial occlusion, camera transitions, severe image blur, incomplete presentation of the AI anchor, or prolonged static imagery were excluded from the candidate pool. All audio tracks were disabled so that voice, pitch, and other auditory cues could not influence visual appearance judgments. The videos averaged 68 s in duration (SD = 4.8; range = 50–86 s). Two uniformly trained research assistants watched and independently rated all 255 videos; no video was evaluated by only one rater. After watching each complete video, each rater completed three seven-point appearance-realism items. The items assessed whether the anchor visually resembled a real person, appeared visually natural, displayed human-like visual characteristics, and showed good internal consistency (Cronbach's $\alpha = 0.851$).

Both raters were blinded to the theoretical hypotheses, the study purpose, and the specific content of H1 and H4. The rating files and accompanying materials did not provide the raters with pitch-variability measures, mean-pitch measures, new-danmaku counts, follower-base data, sales-performance data, or other analytic outcomes. All files were presented using computer-generated anonymous identifiers whose filenames contained no account names, interaction levels, or research-condition information. For each video, the three item scores were first averaged within the rater to form a rater-video appearance-realism score. For each rater, the three video scores belonging to the same anchor were then averaged to form a rater-anchor score. Finally, the two raters' scores for the same anchor were averaged to produce the final anchor-level appearance-realism measure, which was matched to all product-presentation segments associated with that anchor.

Interrater reliability was evaluated across the 255 complete videos after the three items had been combined at the video level. A two-way random-effects, absolute-agreement intraclass correlation for the mean of two raters indicated high agreement, ICC (2, 2) = 0.82, 95% CI [0.75, 0.95], supporting aggregation of the two ratings. Because the final moderator was formed from independent ratings of silent visual material, it captures visual human likeness rather than vocal characteristics.

4.2.4. Control Variables

To reduce potential omitted-variable bias, we controlled for relevant textual, vocal, visual, account-level, and livestream-level variables. First, we controlled for textual linguistic features. Using the LIWC dictionary, we calculated the proportions of function words, cognitive-process words, social words, money-related words, and emotional valence in the transcript of each product presentation segment.

Second, at the vocal level, we controlled for the AI anchor's mean pitch, mean loudness, loudness variability, speech rate, pause ratio, number of voiced speech segments, and presence of background music. Mean pitch, mean loudness, and loudness variability were extracted using openSMILE; speech rate was calculated as the number of transcribed Chinese characters per unit of time; and the number of voiced speech segments was identified and counted using voice activity detection.

Third, at the visual level, we controlled for the AI anchor's positive and negative facial expressions. Specifically, the Face++ facial detection API was used to identify facial emotion categories in video frames. Positive facial expression was measured using the confidence score for "happiness," whereas negative facial expression was measured as the

mean confidence score across "anger," "disgust," "fear," and "sadness." We also controlled for facial detection confidence, BRISQUE image distortion [49], NIMA image quality [50], lip-sync accuracy, and movement fluency to distinguish appearance realism from technical image quality.

Fourth, at the account and livestream levels, we controlled for account follower count, sales reputation, average monthly livestreaming frequency, the number of online viewers immediately before the product presentation, product price, discount rate, promotional intensity, and the frequency of calls to interact. The full model additionally included product-category and day-of-week fixed effects.

4.3. Model Specification

The dependent variable was the number of new danmaku comments in each product presentation segment, a nonnegative integer count variable. Its mean was 28.47, and its variance was 2937.82; the variance substantially exceeded the mean, indicating pronounced overdispersion. We therefore estimated negative binomial regression models [51,52]. In addition, the 4322 product presentation segments were nested within 330 livestreams and 85 accounts, and multiple livestreams and product segments from the same account may share anchor characteristics, account quality, and audience base. To account for within-account dependence, all main models used account-clustered robust standard errors. The model was specified as follows:

$$\begin{aligned}\log E\left(Y_{ijk} \mid \mathrm{X}\right) = \; & \log\left(Duration_{ijk}/60\right) + \beta_0 + \beta_1 PitchVar_{ijk} \\ & + \beta_2 PitchVar_{ijk}^2 + \beta_3 Realism_k \\ & + \beta_4\left(PitchVar_{ijk} \times Realism_k\right) \\ & + \beta_5\left(PitchVar_{ijk}^2 \times Realism_k\right) \\ & + \gamma' Controls_{ijk} + FE_{category} + FE_{weekday}\end{aligned}$$

where i indexes product presentation segments, j indexes livestream sessions, and k indexes AI anchors. Y_{ijk} denotes consumer engagement, measured as the number of new danmaku comments generated during the ith product presentation segment of the jth livestream featuring the kth AI anchor. $\log\left(Duration_{ijk}/60\right)$ is the offset term, which adjusts for differences in segment duration and expresses exposure relative to a 60 s interval. $PitchVar_{ijk}$ denotes the AI anchor's standardized pitch variability during the segment, and $PitchVar_{ijk}^2$ is its quadratic term used to test the proposed inverted U-shaped relationship. $Realism_k$ denotes the appearance realism of the kth AI anchor. $PitchVar_{ijk} \times Realism_k$ and $PitchVar_{ijk}^2 \times Realism_k$ represent the interactions of appearance realism with the linear and quadratic terms of pitch variability, respectively. $Controls_{ijk}$ denotes the vector of text-, audio-, video-, account-, and livestream-level control variables. $FE_{category}$ and $FE_{weekday}$ denote product-category and day-of-week fixed effects, respectively.

4.4. Results

We sequentially estimated four negative binomial regression models. Model 1 included only the control variables. Model 2 added the linear and squared terms of pitch variability. Model 3 further added appearance realism and its interactions with the linear and squared pitch-variability terms. Model 4 added product-category and day-of-week fixed effects to Model 3. All models used account-clustered robust standard errors. Table 2 reports the regression results.

Table 2. Negative binomial regression model results on the impact of AI anchor pitch variability on consumer engagement.

Variable	Model 1	Model 2	Model 3	Model 4
Pitch variability	—	0.224 *** (0.022)	0.251 *** (0.021)	0.254 *** (0.022)
Pitch variability^2	—	−0.188 *** (0.020)	−0.203 *** (0.017)	−0.201 *** (0.017)
Appearance realism	—	—	0.173 ** (0.063)	0.175 ** (0.058)
Pitch variability × appearance realism	—	—	0.057 ** (0.019)	0.056 ** (0.018)
Pitch variability^2 × appearance realism	—	—	−0.082 *** (0.019)	−0.085 *** (0.020)
Control variables	Yes	Yes	Yes	Yes
Product-category fixed effects	No	No	No	Yes
Day-of-week fixed effects	No	No	No	Yes
Account-clustered SEs	Yes	Yes	Yes	Yes
Log likelihood	−17,608.58	−17,477.96	−17,447.21	−17,423.83
AIC	35,271.15	35,013.91	34,958.42	34,947.66
N	4322	4322	4322	4322

Note. Account-clustered robust standard errors are reported in parentheses. *** $p < 0.001$; ** $p < 0.01$.

Model 2 showed that the linear term of pitch variability was significantly positive ($\beta = 0.224$, SE = 0.022, $p < 0.001$), whereas its squared term was significantly negative ($\beta = -0.188$, SE = 0.020, $p < 0.001$). After appearance realism and the interaction terms were added in Model 3, the directions and significance of the linear and squared pitch-variability coefficients remained stable. After further controlling for product-category and day-of-week fixed effects, the linear term remained significantly positive in Model 4 ($\beta = 0.254$, SE = 0.022, $p < 0.001$), and the squared term remained significantly negative ($\beta = -0.201$, SE = 0.017, $p < 0.001$).

However, a significantly positive linear term and a significantly negative squared term alone are insufficient to establish an inverted U-shaped relationship. We therefore examined the conditional slopes of pitch variability at the lower and upper ends of the sample and estimated the curve's turning point and confidence interval. Table 3 presents the results.

Table 3. Boundary Slopes, Conditional Slopes, and Turning Points.

Appearance Realism Level	Slope at P10	95% CI	Slope at P90	95% CI	Turning Point (Semitone SD)
Low realism (−1 SD)	0.499	[0.377, 0.622]	−0.107	[−0.245, 0.031]	3.833
Mean realism	0.769	[0.660, 0.877]	−0.270	[−0.359, −0.180]	3.628
High realism (+1 SD)	1.038	[0.855, 1.221]	−0.432	[−0.560, −0.303]	3.544

Note. The 95% CI for the overall turning point is [3.499, 3.757]. Under low appearance realism, the confidence interval for the slope at P90 includes zero.

At the mean level of appearance realism, the slope of pitch variability at P10 was 0.769 (95% CI [0.660, 0.877]), whereas the slope at P90 was −0.270 (95% CI [−0.359, −0.180]). Thus, the slope was significantly positive at the lower end of the sample and signifi-

cantly negative at the upper end, satisfying the boundary-slope criterion for an inverted U-shaped relationship.

At the mean level of appearance realism, the turning point was 3.628 semitones (95% CI [3.499, 3.757]). This turning point fell within both the observed range of pitch variability (0.800–6.200) and the sample's P5–P95 range (1.482–4.608), indicating that the estimated turning point was not driven entirely by a small number of extreme upper-tail observations. Taken together, these findings support H1.

Regarding moderation, Model 4 showed that the interaction between appearance realism and the linear pitch-variability term was significantly positive ($\beta = 0.056$, SE = 0.018, $p = 0.002$), whereas the interaction with the squared term was significantly negative ($\beta = -0.085$, SE = 0.020, $p < 0.001$). Thus, the curvature of the relationship between pitch variability and consumer engagement became more pronounced as appearance realism increased. Conditional-slope analyses further showed a complete inverted U-shaped pattern at mean and high levels of appearance realism: the slope at P10 was significantly positive, and the slope at P90 was significantly negative. Under low appearance realism, the P90 slope was negative, but its 95% confidence interval included zero; hence, the decline on the right side of the curve was not statistically significant in that condition. These results support H4. Figure 3 illustrates the moderating effect.

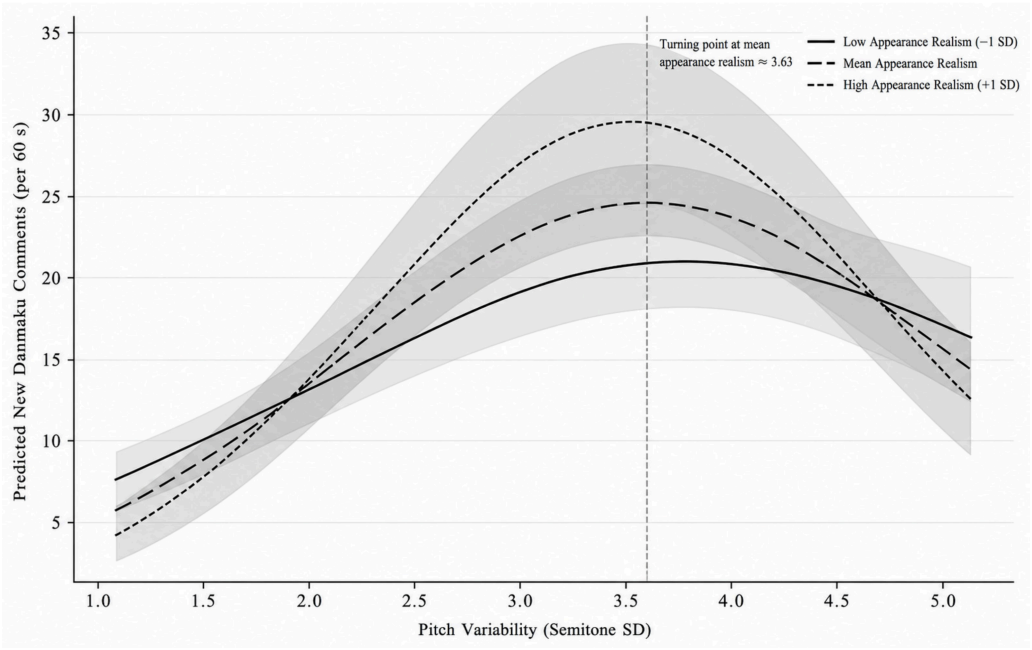

Figure 3. The moderating effect of appearance realism. (Note. The grey shaded areas represent 95% confidence intervals around the predicted values.)

4.5. Robustness Check

To determine whether the conclusions depended on a particular model specification or data-processing choice, we conducted robustness checks addressing five issues: hierarchical dependence, count distribution, variable measurement, outlier treatment, and the zero-count process. Each check altered only one key analytical specification; all other controls, fixed effects, and the segment-duration offset were held consistent with Model 4. Tables 4 and 5 report the results.

Table 4. Robustness Checks for the Inverted U-Shaped Main Effect.

Model Specification	N	Linear Term β	SE	Squared Term β	SE
Main model: NB2, account clustering	4322	0.254 ***	0.022	−0.201 ***	0.017
Livestream-session-clustered SEs	4322	0.254 ***	0.025	−0.201 ***	0.017
Two-way account + livestream clustering	4322	0.254 ***	0.022	−0.201 ***	0.017
PPML, account-clustered SEs	4322	0.274 ***	0.027	−0.200 ***	0.026
Alternative measure: semitone IQR	4322	0.209 ***	0.019	−0.149 ***	0.018
Alternative measure: Hz coefficient of variation	4322	0.154 ***	0.020	−0.100 ***	0.014
Exclude pitch P1/P99	4234	0.253 ***	0.019	−0.191 ***	0.018
Winsorize new danmaku at P99	4322	0.249 ***	0.022	−0.202 ***	0.016
Exclude zero-danmaku segments	3894	0.232 ***	0.021	−0.178 ***	0.015
PPML, duration offset	4322	0.270 ***	0.027	−0.199 ***	0.027
PPML, viewer-minutes offset	4322	0.291 ***	0.029	−0.204 ***	0.027
Account fixed effects, duration offset	4322	0.259 ***	0.028	−0.188 ***	0.025
Account fixed effects, viewer-minutes	4322	0.277 ***	0.030	−0.198 ***	0.024
Session fixed effects, duration offset	4322	0.271 ***	0.025	−0.201 ***	0.025
Session fixed effects, viewer-minutes	4322	0.292 ***	0.027	−0.212 ***	0.024
CRE within-account effect, viewer-minutes	4322	0.246 ***	0.032	−0.195 ***	0.027

Note. *** $p < 0.001$. All continuous focal variables were restandardized within the relevant analytical sample. PPML denotes Poisson pseudo-maximum likelihood; CRE denotes the correlated-random-effects decomposition. Duration offsets are expressed relative to one minute, and viewer-minutes equal mean concurrent viewers multiplied by segment duration in minutes. All specifications include the applicable control variables and fixed effects described in the text; standard errors are clustered by account.

Table 5. Robustness Checks for the Moderating Effect of Appearance Realism.

Model Specification	N	X × W, β (SE)	*p* Value	X^2 × W, β (SE)	*p* Value
Main model: NB2, account clustering	4322	0.056 (0.018)	0.002	−0.085 (0.020)	<0.001
Livestream-session-clustered SEs	4322	0.056 (0.024)	0.019	−0.085 (0.019)	<0.001
Two-way account + livestream clustering	4322	0.056 (0.018)	0.002	−0.085 (0.020)	<0.001
PPML, account-clustered SEs	4322	0.077 (0.028)	0.005	−0.097 (0.026)	<0.001
Alternative measure: semitone IQR	4322	0.029 (0.018)	0.112	−0.068 (0.021)	0.002
Alternative measure: Hz coefficient of variation	4322	−0.020 (0.015)	0.180	−0.023 (0.014)	0.097
Exclude pitch P1/P99	4234	0.058 (0.019)	0.002	−0.069 (0.019)	<0.001
Winsorize new danmaku at P99	4322	0.053 (0.019)	0.004	−0.083 (0.018)	<0.001
Exclude zero-danmaku segments	3894	0.054 (0.018)	0.002	−0.074 (0.018)	<0.001
PPML, duration offset	4322	0.080 (0.028)	0.004	−0.095 (0.026)	<0.001
PPML, viewer-minutes offset	4322	0.081 (0.031)	0.008	−0.096 (0.027)	<0.001
Account fixed effects, duration offset	4322	0.057 (0.027)	0.033	−0.094 (0.024)	<0.001
Account fixed effects, viewer-minutes	4322	0.047 (0.030)	0.125	−0.095 (0.024)	<0.001
Session fixed effects, duration offset	4322	0.062 (0.022)	0.005	−0.103 (0.022)	<0.001
Session fixed effects, viewer-minutes	4322	0.044 (0.026)	0.088	−0.104 (0.022)	<0.001
CRE within-account effect, viewer-minutes	4322	−0.036 (0.038)	0.334	−0.077 (0.027)	0.004

Note. X denotes standardized pitch variability, and W denotes standardized appearance realism. PPML denotes Poisson pseudo-maximum likelihood; CRE denotes the within-account component of the correlated-random-effects decomposition. All specifications include the applicable controls and fixed effects described in the text, with account-clustered standard errors. The quadratic interaction is comparatively stable, whereas the linear interaction is sensitive to pitch operationalization, exposure, and within-account specifications.

First, we addressed hierarchical dependence and unobserved heterogeneity more directly. In addition to the account-clustered NB2 model, we estimated Poisson pseudo-maximum-likelihood count models with account fixed effects and livestream-session fixed effects. We also estimated a correlated-random-effects decomposition that separated within-account from between-account pitch variability. The inverted U-shaped association per-

sisted in every specification. With account fixed effects and the duration offset, the linear and squared coefficients were 0.259 (SE = 0.028, $p < 0.001$) and −0.188 (SE = 0.025, $p < 0.001$), respectively. With session fixed effects, they were 0.271 (SE = 0.025, $p < 0.001$) and −0.201 (SE = 0.025, $p < 0.001$). In the correlated-random-effects model, the within-account linear effect was 0.246 (SE = 0.032, $p < 0.001$) and the within-account squared effect was −0.195 (SE = 0.027, $p < 0.001$). Thus, the nonlinear association was also present in variation occurring within the same account and livestream session.

Second, we examined both the count-distribution assumption and the exposure definition. The main analysis used NB2 regression with an offset for segment duration. PPML models produced the same positive linear and negative quadratic pattern. Because opportunities to comment depend jointly on time and audience size, we additionally defined viewer-minutes as mean concurrent viewers multiplied by segment duration in minutes and used its logarithm as the exposure offset. In the baseline PPML viewer-minutes model, the linear term was 0.291 (SE = 0.029, $p < 0.001$) and the squared term was −0.204 (SE = 0.027, $p < 0.001$). With account fixed effects and viewer-minutes, the corresponding estimates were 0.277 (SE = 0.030, $p < 0.001$) and −0.198 (SE = 0.024, $p < 0.001$); with session fixed effects, they were 0.292 (SE = 0.027, $p < 0.001$) and −0.212 (SE = 0.024, $p < 0.001$). The main nonlinear association, therefore, did not depend on using duration alone as exposure.

Third, regarding variable measurement, we replaced the semitone standard deviation with the semitone interquartile range and the raw-Hz coefficient of variation. Under both alternative measures, the linear term was significantly positive and the squared term significantly negative, further supporting the inverted U-shaped main effect. The moderating effect, however, was sensitive to how pitch variability was operationalized: with the semitone interquartile range, the quadratic interaction remained significant, but the linear interaction did not; with the raw-Hz coefficient of variation, neither interaction reached the 0.05 significance level.

Fourth, regarding outliers, we removed 88 segments with pitch variability below P1 or above P99 to rule out the influence of sparse tail observations on the quadratic estimate. We also winsorized new danmaku counts at P99 to reduce the leverage of extremely high-engagement segments. Under both treatments, the linear and squared pitch-variability terms and both interaction terms retained the directions and significance observed in the main model. In particular, after removing extreme pitch observations, the linear coefficient was 0.253 and the squared coefficient was −0.191, indicating that the inverted U-shaped relationship was not primarily driven by extreme upper-tail observations.

Finally, regarding the zero-count process, 428 segments had zero danmaku comments, representing 9.90% of the sample. The main NB2 model predicted a zero-count rate of 9.58%, only 0.32 percentage points below the observed rate, providing no indication of a substantial excess-zero process. We therefore did not treat a zero-inflated model as necessary for the main analysis. After excluding zero-danmaku segments and re-estimating the model, the linear and squared pitch-variability terms and both interaction terms retained their directions and significance, indicating that the main findings were not driven by a concentration of zero-count observations.

Across all specifications, the linear pitch-variability term remained significantly positive, and the squared term remained significantly negative. The inverted U-shaped main association therefore received robust support across clustering choices, count-distribution assumptions, pitch measures, outlier treatments, zero-count treatments, fixed-effects specifications, within-account decomposition, and alternative exposure definitions. The moderation evidence was less uniform. The quadratic interaction remained negative and significant in the fixed-effects and correlated-random-effects models, but the linear interaction became nonsignificant in the account fixed-effects viewer-minutes model ($\beta = 0.047$,

SE = 0.030, p = 0.125), was marginal in the session fixed-effects viewer-minutes model (β = 0.044, SE = 0.026, p = 0.088), and was nonsignificant in the within-account decomposition (β = −0.036, SE = 0.038, p = 0.334). Accordingly, H1 received robust support, whereas H4 received conditional support that was sensitive to the operationalization of pitch and to the separation of within- and between-account variation.

5. Study 2

Study 1, using real-world AI-anchored livestream-commerce data from Douyin, identified a robust inverted U-shaped association between AI anchor pitch variability and consumer engagement and provided conditional, specification-sensitive evidence concerning the moderating role of appearance realism. Although real-world livestream data offer high external validity, they cannot directly reveal the underlying psychological process. Study 2 therefore uses an experiment to examine whether the proposed pattern occurs under controlled pitch conditions and whether the resulting indirect-effect evidence is consistent with the proposed role of perceived authenticity.

5.1. Experimental Design

Study 2 employed a preregistered mixed experimental design (AsPredicted #303,499). The registration was completed on 24 July 2026, before any data were collected. AI anchor pitch variability was a three-level between-participants factor (low, moderate, and high), whereas AI anchor identity and product varied within participants as stimulus factors. The full stimulus pool followed a 4 × 4 × 3 structure, comprising four AI anchor identities, four products, and three pitch-variability conditions, for a total of 48 video stimuli. The four products were nuts, instant oatmeal, wireless Bluetooth earphones, and a casual hat. Two anchors had feminine appearances, and two had masculine appearances; within each gender presentation, one anchor had relatively high and one relatively low intended appearance realism.

Each participant was assigned to only one pitch-variability condition. Within that condition, the participant watched nine product-introduction videos formed by combinations of three AI anchors and three products. Using a balanced incomplete-block procedure, we randomly selected three of the four anchors and three of the four products for each participant. The selected anchors and products formed a complete 3 × 3 combination, ensuring that each selected anchor presented each selected product. The nine videos were shown in a randomized order across participants. Compared with requiring each participant to watch all 16 anchor–product combinations, this design reduced fatigue, provided relatively balanced exposure for all four anchors and products across the full sample, and prevented the pitch-variability condition from being confounded with any single anchor, product, or stimulus file.

5.2. Stimulus Development

The three pitch-variability conditions used identical AI anchor appearances, product information, product-introduction scripts, subtitles, livestream backgrounds, and video-production procedures. For each anchor-product combination, the low-, moderate-, and high-pitch-variability versions were generated from the same source audio and script. All F0 manipulations were completed in Praat 6.1.35 and batch processed through Python and Parselmouth, which calls Praat's underlying acoustic-analysis and resynthesis routines. Acoustic verification was performed in Python 3.13 using Praat–Parselmouth 0.45 to call the Praat 6.1.35 pitch-analysis functions. F0 was extracted at a 10 ms time step. For all four anchors, pitch floor and pitch ceiling were fixed at 50 and 600 Hz, respectively, and remained unchanged across all stimuli. F0 contours were processed on the semitone scale. For each

sentence, the mean semitone F0 was calculated from valid voiced frames, and deviations around that mean were scaled as $x't = \mu + k(xt - \mu)$, with k = 0.50, 1.00, and 1.50 for the low-, moderate-, and high-variability conditions, respectively. This transformation preserved each sentence's mean semitone F0. Unvoiced frames and pauses were not assigned F0, and interpolation or smoothing was not performed across unvoiced boundaries. Isolated F0-detection errors within continuous voiced regions were corrected; when smoothing was required, a three-frame median filter was applied within, but never across, voiced-region boundaries. The modified PitchTier was returned to the Praat manipulation object and resynthesized using overlap-add. DurationTier, formants, and intensity contours were not actively modified, and the original sampling rate, bit depth, and channel configuration were retained.

Mean F0, F0 standard deviation, F0 range, duration, intensity, and formant measures were recomputed for every output file to verify the realized manipulation and assess unintended acoustic differences, and audiovisual synchronization was checked during video assembly. No stimulus was reprocessed or excluded on the basis of these checks. Pitch variability, operationalized as the within-file standard deviation of semitone F0, averaged 1.88 semitones in the low condition, 3.78 in the moderate condition, and 5.96 in the high condition. Mean F0 was 164.67, 165.36, and 164.77 Hz, respectively; mean loudness was −16.47, −16.48, and −16.50 LUFS; and mean speech rate was 4.02, 4.02, and 3.98 Chinese characters per second. Mean pause ratios ranged from 0.137 to 0.144, and mean video durations ranged from 59.61 to 60.26 s. These objective checks indicate that the intended separation in pitch variability was achieved while the reported non-target acoustic characteristics remained closely aligned.

Appearance realism was independently evaluated using a three-item, seven-point Likert measure. The rating materials comprised the 16 complete base videos formed by four AI anchors presenting four products; no shorter excerpts were sampled, and the audio track was disabled during rating. Two uniformly trained research assistants independently viewed and rated all 16 videos. They were blinded to the study hypotheses and purposes, H1 and H4, the pitch-variability measures, danmaku outcomes, and other account-performance information. For each video and rater, the three items were first averaged to form a rater-video score. For each rater, the four product-video scores associated with the same anchor were then averaged to form an anchor-level score, after which the two raters' anchor-level scores were averaged to create the independent anchor-level appearance-realism measure. Internal consistency was satisfactory (Cronbach's $\alpha = 0.892$). Interrater reliability, assessed across the 16 videos using a two-way random-effects, absolute-agreement ICC for the mean of two raters, was high, ICC (2, 2) = 0.87, 95% CI [0.74, 0.92].

5.3. Participants and Procedure

On 25 July 2026, we obtained 300 submissions through Credamo [53], with 100 participants randomized to each pitch-variability condition. In accordance with the prespecified exclusion rules, 12 participants were excluded: four because required questions were left unanswered and eight because they completed the questionnaire in less than 30 s. The exclusions comprised three participants in the low condition, five in the moderate condition, and four in the high condition. No participant was excluded under the other prespecified criteria. The final confirmatory sample comprised 288 participants (Mage = 31.14 years, SD = 6.83; 62.85% women): 97 in the low condition, 95 in the moderate condition, and 96 in the high condition. These participants contributed 2592 retained trial records before model-specific missing-data deletion.

At the start of the experiment, participants provided electronic informed consent and were informed that participation was voluntary, that they could withdraw, that their

responses would be treated anonymously, and that the data would be used for academic research. The experiment was approved by the institutional ethics committee of the authors' institution (Approval No. NAU202607126), and each participant received RMB 5 upon completion. Participants were then asked to imagine browsing a livestream room in which an AI anchor introduced products, randomly assigned to one pitch-variability condition, and shown nine product-introduction videos in randomized order. After each video, they completed the focal measures and manipulation-purity assessments; demographic questions were administered at the end.

5.4. Measures and Assessment of Reliability and Validity

To ensure measurement reliability and validity, the focal variables were measured using established scales adapted to the AI-anchored livestream context. Perceived pitch variability was adapted from the two-item scale of Hildebrand et al. (2020) [13], and AI anchor familiarity and livestream shopping experience were measured with three-item scales based on Park et al. (1994) [54]. The appearance-realism items were developed for the present context with reference to Miller et al.'s (2023) conceptualization of perceived human likeness [55]. Voice naturalness used a three-item scale based on Nussbaum et al. (2025) [38], emotional intensity was adapted from Holz et al. (2021) [56], and perceived vocal exaggeration and semantic–prosodic fit were developed from Hildebrand et al.'s (2020) voice-perception framework [13]. Standardized loadings for multi-item constructs were at least 0.703; Cronbach's α ranged from 0.750 to 0.904; all composite reliability values exceeded 0.70; and all average variance extracted values exceeded 0.50. These results indicate good reliability and validity. Table 6 reports the items and psychometric properties.

Table 6. Scale Items and Reliability and Validity Assessment.

Construct	Item	Loading	α	CR	AVE
Perceived pitch variability [13]	The AI anchor's pitch varies substantially while speaking.	0.905	0.900	0.900	0.818
	The AI anchor's voice shows pronounced rises and falls in pitch.	0.905			
Perceived authenticity [57]	This AI anchor's expression is sincere.	0.771	0.820	0.820	0.602
	This AI anchor comes across as authentic and credible.	0.777			
	This AI anchor's presentation is unaffected.	0.781			
Consumer engagement intention [19]	I would be willing to post danmaku comments or other comments in this livestream room.	0.747	0.809	0.809	0.586
	I would be willing to like this livestream content.	0.778			
	I would be willing to continue watching or revisit this AI anchor's livestream room.	0.771			
Appearance realism [55]	This AI anchor visually resembles a real human.	0.841	0.892	0.885	0.719
	This AI anchor's visual presentation looks natural.	0.846			
	This AI anchor has clearly human-like visual features.	0.856			
Voice naturalness [38]	The AI anchor's voice sounds natural.	0.802	0.838	0.838	0.633
	The AI anchor's voice sounds like a real person speaking.	0.788			
	The AI anchor's voice does not sound mechanical.	0.797			
Emotional intensity [56]	The AI anchor's voice conveys strong emotion.	0.703	0.750	0.752	0.502
	The AI anchor expresses emotion forcefully while speaking.	0.718			
	The emotional arousal in the AI anchor's voice is high.	0.705			

Table 6. *Cont.*

Construct	Item	Loading	α	CR	AVE
AI anchor familiarity [54]	I am familiar with AI anchors.	0.898	0.904	0.905	0.760
	I understand how AI anchors conduct livestream product presentations.	0.869			
	Compared with the average consumer, I am relatively familiar with AI anchors.	0.849			
Livestream shopping experience [54]	I am familiar with the livestream shopping process.	0.773	0.829	0.831	0.620
	I have substantial livestream shopping experience.	0.819			
	Compared with the average consumer, I know more about livestream shopping.	0.770			
Perceived vocal exaggeration [13]	The AI anchor's intonation sounds exaggerated.	1.000	—	—	—
Semantic–prosodic fit [13]	The emphasis and pitch contours in the AI anchor's voice match the content being delivered.	1.000	—	—	—
Information sufficiency [58]	The video provides sufficient product information.	1.000	—	—	—
Product attractiveness [59]	The product shown in the video is attractive to me.	1.000	—	—	—
Audiovisual quality	The video has good visual and audio quality.	1.000	—	—	—

5.5. Analytical Strategy

Because each participant evaluated nine videos, the preregistered primary regressions used participant-clustered robust standard errors and included AI anchor fixed effects, product fixed effects, standardized trial order, AI anchor familiarity, and livestream-shopping experience. To make the coefficients equal the intended group comparisons, the planned contrasts were rescaled without changing their hypothesis tests. C1 directly compared the moderate condition with the average of the low and high conditions (low = $-1/3$, moderate = $2/3$, high = $-1/3$), whereas C2 directly compared high with low variability (low = $-1/2$, moderate = 0, high = $1/2$). Scale scores were calculated only when all preregistered items were available, and no missing item was imputed. The engagement-intention model used 2506 trials, the perceived-authenticity model used 2482 trials, and 2399 trials contained complete mediator and outcome scores. Missingness was summarized by condition, anchor, product, and trial order and modeled using participant-clustered logistic regressions. Sensitivity was assessed through analyses aggregating all available trials for every participant and through inverse-probability weighting of complete mediation trials.

The indirect-effect analysis treated the design as a 2-1-1 multilevel structure because the pitch condition varied between participants, whereas perceived authenticity and engagement intention were repeatedly measured within participants. The primary indirect-effect specification aggregated all available trial responses to the participant level, included indicators for exposure to the included anchors and products, and retained all 288 participants. A complementary trial-level model decomposed perceived authenticity into each participant's mean and the trial-specific deviation from that mean, thereby separating between-participant and within-participant associations. Indirect effects were evaluated using 5000 participant-level bootstrap samples; all observations belonging to a sampled

participant were retained together. To address crossed stimulus dependence, mixed models included participant and anchor-product random intercepts, and an additional specification allowed the C1 and C2 treatment effects to vary randomly across the 16 anchor-product combinations. The preregistration specified the hypotheses, target sample size, exclusion criteria, focal outcomes, planned contrasts, and participant-clustered regression models. The rescaling of the contrasts, multilevel mediation decomposition, missing-data sensitivity analyses, and stimulus random-slope model were conducted as additional analyses to improve coefficient interpretation and address the multilevel and crossed structures of the data.

5.6. Results

We first assessed whether the pitch-variability manipulation was effective. Because pitch variability was manipulated between participants and each participant evaluated nine videos, we averaged perceived pitch variability across the nine trials for each participant and conducted a one-way ANOVA. Perceived pitch variability differed significantly across the three conditions, $F(2, 285) = 455.62$, $p < 0.001$, and increased monotonically from the low condition ($M = 2.49$, $SD = 0.86$) to the moderate condition ($M = 4.69$, $SD = 0.77$) and the high condition ($M = 6.01$, $SD = 0.81$). Post hoc comparisons showed significant differences between every pair of conditions (all $p < 0.001$). Participants therefore clearly distinguished the three pitch-variability levels, confirming the intended manipulation.

To examine whether the pitch manipulation was accompanied by statistically detectable differences in other voice perceptions or stimulus evaluations, we analyzed voice naturalness, emotional intensity, perceived vocal exaggeration, semantic-prosodic fit, information sufficiency, product attractiveness, and audiovisual quality. The models used participant-clustered robust standard errors and included AI anchor fixed effects, product fixed effects, trial order, AI anchor familiarity, and livestream-shopping experience. As shown in Table 7, neither corrected C1 nor corrected C2 reached statistical significance for any of these measures (all $p \geq 0.078$). These results indicate that no statistically detectable condition differences were observed on the measured alternative explanations; they should not be interpreted as evidence of equivalence. The confidence intervals in Table 7 show the range of effects that remains compatible with the data.

Table 7. Manipulation-Purity and Alternative-Explanation Tests.

Variable	Contrast	β	SE	*p*	95% CI
Voice naturalness	C1: Moderate vs. low/high	0.025	0.039	0.518	[−0.051, 0.101]
	C2: High vs. low	−0.056	0.040	0.165	[−0.135, 0.023]
Emotional intensity	C1: Moderate vs. low/high	−0.018	0.034	0.596	[−0.085, 0.049]
	C2: High vs. low	−0.016	0.040	0.682	[−0.094, 0.061]
Perceived vocal exaggeration	C1: Moderate vs. low/high	−0.059	0.036	0.101	[−0.130, 0.011]
	C2: High vs. low	−0.075	0.042	0.078	[−0.158, 0.008]
Semantic–prosodic fit	C1: Moderate vs. low/high	0.015	0.032	0.638	[−0.047, 0.077]
	C2: High vs. low	−0.008	0.041	0.842	[−0.089, 0.072]
Information sufficiency	C1: Moderate vs. low/high	−0.015	0.032	0.634	[−0.077, 0.047]
	C2: High vs. low	0.001	0.037	0.989	[−0.071, 0.072]
Product attractiveness	C1: Moderate vs. low/high	−0.017	0.039	0.659	[−0.094, 0.059]
	C2: High vs. low	−0.040	0.042	0.343	[−0.123, 0.043]
Audiovisual quality	C1: Moderate vs. low/high	0.031	0.029	0.275	[−0.025, 0.088]
	C2: High vs. low	0.046	0.032	0.153	[−0.017, 0.109]

We next tested the effect of AI anchor pitch variability on perceived authenticity. A participant-level one-way ANOVA showed a significant effect, $F(2, 285) = 261.21$, $p < 0.001$. Perceived authenticity was highest in the moderate condition (M = 5.41, SD = 0.35) and lower in the low (M = 4.33, SD = 0.39) and high (M = 4.32, SD = 0.40) conditions. Post hoc comparisons showed that the moderate condition exceeded both endpoint conditions (both $p < 0.001$), whereas the low and high conditions did not differ ($p = 0.907$).

With participant-clustered robust standard errors and controls for AI anchor fixed effects, product fixed effects, trial order, AI anchor familiarity, and livestream-shopping experience, corrected C1 significantly predicted perceived authenticity ($\beta = 1.086$, SE = 0.045, $p < 0.001$, 95% CI [0.999, 1.173]), whereas corrected C2 was nonsignificant ($\beta = -0.004$, SE = 0.057, $p = 0.946$, 95% CI [−0.115, 0.108]). Because corrected C1 now directly represents the moderate condition minus the mean of the endpoint conditions, its magnitude can be interpreted as an approximately 1.09-point authenticity advantage for moderate pitch variability. These results support H2 within the tested conditions.

Consumer engagement intention showed the same descriptive pattern. A participant-level one-way ANOVA revealed a significant effect of pitch variability, $F(2, 285) = 123.52$, $p < 0.001$. Engagement intention was highest in the moderate condition (M = 5.03, SD = 0.46), exceeding both the low (M = 4.09, SD = 0.48) and high (M = 4.16, SD = 0.44) conditions (both $p < 0.001$); the endpoint conditions did not differ ($p = 0.317$). In the participant-clustered model, corrected C1 was significant ($\beta = 0.908$, SE = 0.057, $p < 0.001$, 95% CI [0.796, 1.020]), whereas corrected C2 was not ($\beta = 0.064$, SE = 0.066, $p = 0.330$, 95% CI [−0.065, 0.193]). Thus, the moderate condition exceeded the average of the endpoint conditions by approximately 0.91 points, supporting H1 within the experimental range.

The focal results were unchanged when all 300 randomized participants were analyzed. Corrected C1 remained significant for engagement intention ($\beta = 0.909$, SE = 0.055, $p < 0.001$, 95% CI [0.801, 1.017]) and perceived authenticity ($\beta = 1.085$, SE = 0.043, $p < 0.001$, 95% CI [1.000, 1.170]); corrected C2 remained nonsignificant for both outcomes. The exclusions, therefore, did not alter the substantive conclusions.

The stimulus random-slope model converged normally. After accounting for crossed participant and anchor-product intercepts and allowing C1 and C2 to vary across the 16 anchor-product combinations, the average corrected C1 effect remained positive and significant ($\beta = 0.908$, SE = 0.104, $p < 0.001$, 95% CI [0.703, 1.112]), whereas corrected C2 remained nonsignificant ($\beta = 0.068$, SE = 0.065, $p = 0.297$, 95% CI [−0.060, 0.196]). The estimated C1 random-slope variance was 0.123, indicating variation in effect magnitude across combinations. Descriptive C1 estimates were positive for all 16 included combinations and ranged from 0.381 to 1.491. This evidence supports directional consistency within the four anchors, four products, and 16 combinations included in the experiment, but the small number of stimulus units does not establish population-level stimulus generalization.

Item-level missingness among retained participants arose when individual questions were left unanswered; it was not generated through planned random item omission. Of the 2592 retained trials, the complete multi-item perceived-authenticity score was unavailable in 110 trials and the engagement-intention score in 86 trials; 2399 trials (92.6%) had complete mediator and outcome scores. Joint missingness rates were 8.0% in the low condition, 7.0% in the moderate condition, and 7.3% in the high condition. Participant-clustered logistic models showed that neither C1 nor C2 predicted missingness in perceived authenticity, engagement intention, or joint mediation completeness. Joint missingness also was not significantly associated with trial order ($p = 0.458$). Perceived-authenticity missingness increased modestly with trial order (OR = 1.113, $p = 0.004$), but every included participant contributed sufficient available responses to calculate participant-level mean authenticity and engagement scores. The all-participant aggregate analysis and the inverse-probability-

weighted complete-trial analysis produced substantively consistent C1 and authenticity effects, reducing concern that the primary pattern was created by complete-case selection.

We next examined the indirect-effect pattern involving perceived authenticity using the 2-1-1 strategy described above. In the participant-level aggregate model, corrected C1 and corrected C2 were entered simultaneously, perceived authenticity was the proposed mediator, and consumer engagement intention was the outcome. All 288 participants contributed aggregate scores based on their available trials. The models controlled for AI anchor familiarity, livestream-shopping experience, and indicators of exposure to the included anchors and products. Five thousand bootstrap samples were generated by resampling participants with replacement.

As shown in Table 8, corrected C1 significantly predicted participant-mean perceived authenticity (a = 1.084, bootstrap SE = 0.047, 95% CI [0.992, 1.177]), whereas corrected C2 did not (a = −0.001, bootstrap SE = 0.056, 95% CI [−0.111, 0.110]). Participant-mean perceived authenticity positively predicted engagement intention at the between-participant level (b = 0.453, bootstrap SE = 0.070, 95% CI [0.312, 0.586]). The complementary trial-level mixed model yielded a similar between-participant authenticity coefficient (β = 0.444, SE = 0.066, $p < 0.001$, 95% CI [0.314, 0.573]) and a significant within-participant association (β = 0.315, SE = 0.024, $p < 0.001$, 95% CI [0.267, 0.363]). This separation avoids combining between-participant treatment effects with undifferentiated trial-level mediator covariation.

Table 8. Multilevel Indirect-Effect Results.

Effect or Path	Estimate	SE	95% CI	*p*
C1 total effect	0.900	0.058	[0.787, 1.012]	<0.001
C1 direct effect	0.410	0.095	[0.229, 0.596]	<0.001
C1 indirect effect through participant-mean authenticity	0.491	0.076	[0.339, 0.635]	—
C2 total effect	0.069	0.066	[−0.057, 0.196]	0.299
C2 direct effect	0.069	0.060	[−0.045, 0.184]	0.250
C2 indirect effect through participant-mean authenticity	−0.000	0.026	[−0.052, 0.049]	—
Authenticity: between-participant association	0.444	0.066	[0.314, 0.573]	<0.001
Authenticity: within-participant association	0.315	0.024	[0.267, 0.363]	<0.001

Note. Corrected C1 directly compares moderate pitch variability with the mean of the low and high conditions; corrected C2 directly compares high with low variability. Participant-level aggregate models include all 288 participants and control for AI anchor familiarity, livestream-shopping experience, and exposure to the included anchors and products. Indirect-effect confidence intervals are based on 5000 participant-level bootstrap samples. The between- and within-participant authenticity coefficients are from the complementary trial-level mixed model with participant and anchor-product random intercepts.

The corrected C1 indirect effect through participant-mean perceived authenticity was 0.491 (bootstrap SE = 0.076, 95% percentile CI [0.339, 0.635]). Its direct effect was 0.410 (bootstrap SE = 0.095, 95% CI [0.229, 0.596]), and its total effect was 0.900 (bootstrap SE = 0.058, 95% CI [0.787, 1.012]). The corrected C2 indirect effect was −0.000 (bootstrap SE = 0.026, 95% CI [−0.052, 0.049]), with a direct effect of 0.069 (95% CI [−0.045, 0.184]) and a total effect of 0.069 (95% CI [−0.057, 0.196]). Thus, the multilevel indirect-effect pattern is consistent with H3, but it remains associational because perceived authenticity was not independently manipulated and causal temporal ordering was not established.

Finally, we examined moderation using the independently rated anchor-level appearance-realism scores described in Section 5.2 rather than treating participants' post-exposure evaluations as an exogenous moderator. With anchor and product fixed effects and participant-clustered inference, the corrected C1 × independent appearance-realism interaction was positive (β = 0.314, SE = 0.031, $p < 0.001$, 95% CI [0.252, 0.375]), whereas the corrected C2 interaction was not significant at the conventional level (β = 0.064, SE = 0.036,

p = 0.075, 95% CI [−0.006, 0.135]). The conditional C1 effects were 0.595 (SE = 0.067, 95% CI [0.463, 0.726]) at one standard deviation below the appearance-realism mean, 0.908 (SE = 0.057, 95% CI [0.796, 1.021]) at the mean, and 1.222 (SE = 0.063, 95% CI [1.098, 1.346]) at one standard deviation above the mean. The predicted moderate-condition advantage therefore increased with independently rated appearance realism within the four anchors studied. Because appearance realism varied across only four anchors and was not experimentally randomized, this result is interpreted as supportive but stimulus-limited evidence rather than a broadly generalizable causal moderation effect.

6. Conclusions and Discussion

As generative artificial intelligence and digital human technologies become increasingly embedded in livestream commerce, AI anchors have gradually shifted from auxiliary display tools to marketing agents that directly interact with consumers [18]. Compared with human streamers, AI anchors have advantages such as stable online availability and controllable expression. However, their nonhuman nature also makes consumers more attentive to whether their expressions appear natural, authentic, and trustworthy [17]. Against this background, this study focuses on the dynamic vocal features of AI anchors, examines how pitch variability influences consumer engagement, and further examines a potential psychological pathway and boundary condition.

Using real-world livestream data and a preregistered multi-stimulus experiment, this research tested complementary components of the proposed framework. Study 1 provided robust evidence of an inverted U-shaped association between pitch variability and observed danmaku engagement, including specifications relying on within-account and within-session variation and viewer-minutes exposure. Study 2 showed that moderate pitch variability produced greater perceived authenticity and engagement intentions than the average of the two endpoint conditions within the four anchors and four products examined. The 2-1-1 indirect-effect pattern was consistent with the proposed role of perceived authenticity but did not establish causal mediation. Evidence concerning appearance realism remained more conditional: Study 1 interactions were specification-sensitive, while the Study 2 interaction used independent ratings but was based on only four anchors.

These findings suggest that AI anchor voice design should not simply aim to maximize vocal variation. One possible interpretation is that moderate pitch variability may make an AI anchor's delivery appear less mechanical and closer to natural interpersonal communication, whereas highly variable delivery might be interpreted as overperformed or artificial. However, Study 2 did not detect statistically significant differences in voice naturalness, perceived vocal exaggeration, or semantic-prosodic fit. These explanations should therefore be treated as theoretically plausible interpretations rather than demonstrated mechanisms. What the experiment directly shows is that moderate variability increased perceived authenticity and engagement intention relative to the average of the endpoint conditions.

This study also contributes to research on AI anthropomorphism and virtual influencers. Existing studies have mainly examined AI marketing agents from the perspectives of appearance, identity disclosure, anthropomorphic image, and virtual influencer authenticity [14,27,45]. These studies often emphasize how visual attractiveness or virtual identity influences consumer trust and attitudes. In contrast, this study shows that vocal features are also important cues through which consumers judge whether an AI anchor appears authentic, natural, and interactive. This finding complements prior research that has emphasized visual cues in AI marketing and highlights the independent value of dynamic vocal cues in AI-enabled livestream marketing.

6.1. Theoretical Contributions

This study makes several theoretical contributions. First, it extends the theoretical boundary of research on streamer vocal features in livestream commerce. Existing livestream commerce research has mainly examined the language content, facial expressions, emotional expression, and interaction strategies of human streamers. Studies on voice have also focused largely on human streamers or salespeople, examining features such as pitch, loudness, and speech rate [12,13]. This study shifts the research context to AI anchors and emphasizes the role of dynamic vocal features in nonhuman marketing agents. The findings show that AI anchor pitch variability significantly influences consumer engagement and that this effect follows an inverted U-shaped pattern. This suggests that voice remains a key social cue in AI-mediated livestream interaction, but its function differs from that of human streamers. For human streamers, pitch variability is often interpreted as emotional involvement or emotional contagion. For AI anchors, however, pitch variability functions more as a designed vocal anthropomorphic cue that shapes consumers' perceptions of naturalness and authenticity.

Second, this study extends AI marketing research from visual realism to vocal realism and dynamic vocal cues. Existing research on AI anchors, virtual humans, and virtual influencers has paid considerable attention to visual attractiveness, identity disclosure, and the degree of anthropomorphism [27,45], while giving relatively limited attention to AI anchors as "voiced agents." This study argues that voice is an important cue in human cognition and social judgment and that it is also a key basis for consumers to judge whether an AI anchor appears authentic and natural. By introducing pitch variability, this study enriches the multimodal perspective on AI marketing agents.

Third, this study provides indirect-effect evidence consistent with the proposed role of perceived authenticity in linking AI-anchored pitch variability to engagement intention. Authenticity research emphasizes that consumers form attitudes and behavioral responses based on whether a marketing agent appears natural, sincere, and trustworthy. The present findings show that moderate pitch variability was associated with stronger authenticity judgments and that these judgments statistically accounted for part of the experimental contrast in engagement intention. Because authenticity was not independently manipulated and the mediator and outcome were assessed within the same post-exposure sequence, this contribution concerns a plausible psychological pathway rather than demonstrated causal mediation.

6.2. Managerial Implications

The findings offer several practical implications. First, brands and platforms should evaluate AI-anchored voices across multiple levels of pitch variability rather than assuming that greater expressiveness is uniformly beneficial. Within the tested range, moderate variability produced higher perceived authenticity and engagement intention than the average of low and high variability. A low-variability voice might sometimes be experienced as mechanical or script-like, but the experiment did not detect a corresponding condition difference in perceived voice naturalness. Firms should therefore base deployment decisions on direct audience testing of authenticity and engagement rather than treating mechanicality as an established explanation.

Second, AI anchor voice design should not automatically equate high variability with more effective emotional expression. High variability did not outperform low variability in the experiment, but participants also did not report a statistically significant increase in vocal exaggeration. Accordingly, interpretations such as artificial or "fake" enthusiasm remain possible explanations rather than observed psychological mechanisms. A prudent

design process would compare multiple voice settings while monitoring authenticity, engagement, semantic-prosodic fit, and perceived exaggeration separately.

Third, firms may consider visual-vocal alignment when designing AI anchors, but this implication remains provisional. Study 1 moderation estimates were sensitive to pitch measurement, exposure, and within-account specifications. Study 2 used independent appearance-realism ratings and found that the moderate-condition advantage increased with realism, but the estimate was based on only four anchors. Firms should therefore test candidate voice-appearance combinations with target audiences rather than assume that one alignment rule will generalize across anchors, products, and platforms.

6.3. Limitations

This study has several limitations that provide concrete directions for future research. First, Study 1 is observational and based on AI-anchored livestream-commerce data from a single platform. Account and livestream-session fixed effects, a correlated-random-effects decomposition, extensive controls, alternative count models, and viewer-minutes exposure substantially reduce—but cannot eliminate—endogeneity from time-varying scripts, promotions, audience composition, and other unobserved conditions. The estimates should therefore be interpreted as robust within-account and within-session associations rather than definitive causal effects.

Second, Study 2 obtained independent appearance-realism ratings from blinded research assistants and used those anchor-level scores in the primary moderation analysis, reducing the post-treatment concern associated with participant ratings. Nevertheless, the moderator varied across only four anchors and was not experimentally randomized. The moderation result therefore remains stimulus-limited and does not establish a population-level causal interaction. Because only two research assistants provided the independent ratings, the high ICC does not completely eliminate the possibility of rater-specific judgment. Future studies should use a larger panel of independent raters, rate a larger set of silent visual-only anchor materials, and experimentally vary visual realism while holding identity and production quality constant.

Third, although Study 2 used four anchors, four products, 48 video files, and random treatment slopes across 16 anchor-product combinations, the C1 effect varied across stimuli, and the number of stimulus units remained small. The results establish directional consistency within the included anchors and products, not population-level generalization. The experiment also compared only three predefined pitch-variability levels. Future preregistered studies should include more anchors, products, scripts, and independently produced exemplars per cell.

Fourth, the 2-1-1 multilevel analysis separated between-participant and within-participant authenticity associations, and the indirect-effect result was robust to all-participant aggregation and inverse-probability-weighted sensitivity analysis. Nevertheless, perceived authenticity and engagement intention were assessed in the same post-exposure sequence. The analysis is therefore consistent with the proposed pathway but does not establish temporal ordering or causal mediation. Future research should manipulate authenticity independently, separate mediator and outcome measurement over time, or use experimental-causal-chain designs.

Fifth, stimulus quality control relied on recomputed acoustic indicators and audiovisual-synchronization checks; no formal human listening test was conducted before the experiment to assess subtle resynthesis artifacts, naturalness, or semantic-prosodic appropriateness. Although the participant measures did not reveal statistically detectable condition differences in voice naturalness, vocal exaggeration, or semantic-prosodic fit, future studies

should add blinded human listening panels and prespecified artifact-rating criteria before fielding the stimuli.

Sixth, iFlytek Tingjian was used to generate Chinese transcripts for the speech-rate and text-based controls, but transcription accuracy was not assessed against a manually annotated subsample. Recognition errors may therefore introduce measurement error, particularly in segments with poor audio quality or atypical pronunciation. Future studies should report character error rates based on human-verified transcripts and examine whether results are robust to alternative transcription systems.

Finally, consumers may differ in their familiarity with digital humans, trust in AI, and expectations of livestream communication. Future research should explicitly examine age, generational cohort, technology readiness, prior AI experience, and cultural background as potential moderators. Such analyses would clarify which audience segments are most receptive to AI anchors and would help firms develop voice designs that are responsive to heterogeneous consumer expectations rather than assuming a single design is appropriate for all users.

Author Contributions: Conceptualization, X.L. and Y.J.; methodology, X.L. and Q.Y.; software, Q.Y.; validation, X.L., Q.Y. and Y.J.; formal analysis, Q.Y.; investigation, X.L. and Q.Y.; resources, X.L. and Y.J.; data curation, Q.Y.; writing—original draft preparation, X.L. and Q.Y.; writing—review and editing, X.L., Q.Y. and Y.J.; visualization, Q.Y.; supervision, Y.J.; project administration, X.L. and Y.J.; funding acquisition, Y.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Liaoning Provincial Social Science Youth Project (grant number LJ112410154046) and the Liaoning University of Technology Doctoral Startup Fund (grant number XB2024012), and was sponsored by the Qing Lan Project of Jiangsu Higher Education Institutions.

Institutional Review Board Statement: This study includes two phases of research. The initial study was approved by the Institutional Ethics Committee of Nanjing Audit University on 13 June 2026 (protocol code NAU202606008). Following the first round of peer review, Study 2 was substantially redesigned, and a new ethical approval for the revised experiment was obtained on 21 July 2026 (protocol code NAU202607126). All new data collection for the revised experiment was conducted after the second approval was granted. The study was conducted in accordance with the Declaration of Helsinki.

Informed Consent Statement: Electronic informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are not publicly available due to privacy and ethical restrictions concerning participant data but are available from the corresponding author on reasonable request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Peltier, J.W.; Dahl, A.J.; Schibrowsky, J.A. Artificial Intelligence in Interactive Marketing: A Conceptual Framework and Research Agenda. *J. Res. Interact. Mark.* **2024**, *18*, 54–90. [CrossRef]
2. Wang, C.L. Editorial—What Is an Interactive Marketing Perspective and What Are Emerging Research Areas? *J. Res. Interact. Mark.* **2024**, *18*, 161–165. [CrossRef]
3. Wang, C.L. Editorial: Demonstrating Contributions through Storytelling. *J. Res. Interact. Mark.* **2025**, *19*, 1–4. [CrossRef]
4. Bai, S.; Yang, Y.; Yu, D.; Tan, Y. Unlocking the Secrets of User Engagement: The Role of Multimodal Information and Sentiment Signals in AI Agent Design. *J. Res. Interact. Mark.* **2026**, *20*, 729–747. [CrossRef]
5. Yang, Q.; Han, Q.; Li, J.; Zhu, J.; Jiang, Y. Exploring the Impact of Certainty Language Style in Social Media Influencers-Generated Content on Consumers' Purchase Behavior. *J. Res. Interact. Mark.* **2026**, *20*, 215–231. [CrossRef]

6. Zheng, X.; Cui, C.; Zhang, C.; Li, D. Who Says What? How Message Appeals Shape Virtual- versus Human-Influencers' Impact on Consumer Engagement. *J. Res. Interact. Mark.* **2026**, *20*, 199–214. [CrossRef]
7. Jayasingh, S.; Sivakumar, A.; Vanathaiyan, A.A. Artificial Intelligence Influencers' Credibility Effect on Consumer Engagement and Purchase Intention. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 17. [CrossRef]
8. Bai, S.; Gui, Z.; Zhou, J.; Zhao, J. Can Consumers Still Form Proximal Sensory Perception in Virtual Anchor Live Streaming? The Impact of the Fit Between Sensory Language Description and Product Attributes. *J. Theor. Appl. Electron. Commer. Res.* **2026**, *21*, 92. [CrossRef]
9. Wang, C.L. Editorial: The Changing Landscape of Marketing Research in the AI Era: Prospects and Challenges. *J. Res. Interact. Mark.* **2026**, *20*, 1–10. [CrossRef]
10. Carew, B.; Peltier, J.; Dahl, A. AI Strategic Orientation and the B2B Social Media Brand Meaning Process: Antecedents, Consequences, and Outcomes. *J. Appl. Bus. Behav. Sci.* **2025**, *1*, 184–209. [CrossRef] [PubMed]
11. Wang, X.; Lu, S.; Li, X.I.; Khamitov, M.; Bendle, N. Audio Mining: The Role of Vocal Tone in Persuasion. *J. Consum. Res.* **2021**, *48*, 189–211. [CrossRef]
12. Luo, L.; Liu, L.; Zheng, Y.; Wang, Y. The Power of Voice: Investigating the Effects of Streamer Voice Characteristics on Sales Performance in Live Streaming E-Commerce. *J. Retail. Consum. Serv.* **2025**, *84*, 104260. [CrossRef]
13. Hildebrand, C.; Efthymiou, F.; Busquet, F.; Hampton, W.H.; Hoffman, D.L.; Novak, T.P. Voice Analytics in Business Research: Conceptual Foundations, Acoustic Feature Extraction, and Applications. *J. Bus. Res.* **2020**, *121*, 364–374. [CrossRef]
14. Sands, S.; Campbell, C.L.; Plangger, K.; Ferraro, C. Unreal Influence: Leveraging AI in Influencer Marketing. *Eur. J. Mark.* **2022**, *56*, 1721–1747. [CrossRef]
15. Yin, Y.; Xu, S. When Virtual Anchors Eclipse the Product: How Form Realism Draws Attention Away but Increases Purchase Intention in Live-Streaming Commerce. *J. Res. Interact. Mark.* **2026**, *20*, 830–846. [CrossRef]
16. Yang, F.; Ying, T.; Liu, X. Echoes of Innovation: Exploring the Use of Voice Assistants to Boost Hotel Reputation. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 46. [CrossRef]
17. Liu, H.; Zhang, P.; Cheng, H.; Hasan, N.; Chiong, R. Impact of AI-Generated Virtual Streamer Interaction on Consumer Purchase Intention: A Focus on Social Presence and Perceived Value. *J. Retail. Consum. Serv.* **2025**, *85*, 104290. [CrossRef]
18. Shui, X.; Bian, S.; Zhang, P. How Can AI Virtual Streamers Gain Consumer Trust to Influence Purchase Intention in Live-Streaming E-Commerce? *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 325. [CrossRef]
19. Alhabash, S.; McAlister, A.R.; Lou, C.; Hagerstrom, A. From Clicks to Behaviors: The Mediating Effect of Intentions to Like, Share, and Comment on the Relationship Between Message Evaluations and Offline Behavioral Intentions. *J. Interact. Advert.* **2015**, *15*, 82–96. [CrossRef]
20. Zhang, Y.; Li, K.; Qian, C.; Li, X.; Yuan, Q. How Real-Time Interaction and Sentiment Influence Online Sales? Understanding the Role of Live Streaming Danmaku. *J. Retail. Consum. Serv.* **2024**, *78*, 103793. [CrossRef]
21. Mir, I.A.; Salo, J. Mapping Content-Driven Engagement and Attitudinal Spillover Effect of Influencer Marketing. *J. Res. Interact. Mark.* **2025**, *19*, 408–425. [CrossRef]
22. Luo, L.; Xu, M.; Wan, F.; Long, F. Too Much of a Good Thing: Nonlinear Effect of Streamers' Language Arousal on Viewer Engagement in Live Streaming. *J. Res. Interact. Mark.* **2026**, *20*, 925–943. [CrossRef]
23. Wang, C. Basic but Frequently Overlooked Issues in Manuscript Submissions: Tips from an Editor's Perspective. *J. Appl. Bus. Behav. Sci.* **2025**, *1*, 140–144. [CrossRef]
24. Li, M.; Cheng, M.; Quintal, V. Decoding the Subtleties: Speech Voice Cues and Their Impacts on Viewer In-Consumption Engagement in Travel Live Streaming. *J. Hosp. Tour. Res.* **2026**, *50*, 494–507. [CrossRef]
25. Yoo, J.W.; Park, J.; Park, H. How Can I Trust You If You're Fake? Understanding Human-like Virtual Influencer Credibility and the Role of Textual Social Cues. *J. Res. Interact. Mark.* **2025**, *19*, 730–748. [CrossRef]
26. Li, C.; Huang, F. The Impact of Virtual Streamer Anthropomorphism on Consumer Purchase Intention: Cognitive Trust as a Mediator. *Behav. Sci.* **2024**, *14*, 1228. [CrossRef] [PubMed]
27. Lou, C.; Kiew, S.T.J.; Chen, T.; Lee, T.Y.M.; Ong, J.E.C.; Phua, Z. Authentically Fake? How Consumers Respond to the Influence of Virtual Influencers. *J. Advert.* **2023**, *52*, 540–557. [CrossRef]
28. Audrezet, A.; De Kerviler, G.; Moulard, J.G. Authenticity under Threat: When Social Media Influencers Need to Go beyond Self-Presentation. *J. Bus. Res.* **2020**, *117*, 557–569. [CrossRef]
29. Lin, X.; Xu, Y.; Hwang, C. Beyond Static Images: How Interactivity, Vividness and Realism Shape Consumer Responses toward 3D Fashion Lookbooks. *J. Res. Interact. Mark.* **2026**, *20*, 593–608. [CrossRef]
30. Chen, A.; Zhang, Y.; Liu, Y.; Lu, Y. Be a Good Speaker in Livestream Shopping: A Speech Act Theory Perspective. *Electron. Commer. Res. Appl.* **2023**, *61*, 101301. [CrossRef]
31. Wang, H.-M.; Zhu, Y.-P.; Lee, K.-T. What Signals Are You Sending? How Signal Consistency Influences Consumer Purchase Behavior in Live Streaming Commerce. *J. Theor. Appl. Electron. Commer. Res.* **2025**, *20*, 109. [CrossRef]

32. Ma, X.; Ren, J.; Lang, X.; Yang, Z.; Li, T. The Influence of Live Video Hosts' Speech Act on Purchase Behaviour. *J. Retail. Consum. Serv.* **2024**, *81*, 103984. [CrossRef]
33. Luo, H.; Cheng, S.; Zhou, W.; Yu, S.; Lin, X. A Study on the Impact of Linguistic Persuasive Styles on the Sales Volume of Live Streaming Products in Social E-Commerce Environment. *Mathematics* **2021**, *9*, 1576. [CrossRef]
34. Ma, X.; Ren, J.; Khor, X.; Wang, R.; Li, T.; Lang, X. Influencing Mechanisms of Live Streamer's Language Strategies on Product Sales. *J. Retail. Consum. Serv.* **2025**, *85*, 104291. [CrossRef]
35. Reeves, B.; Nass, C. *The Media Equation: How People Treat Computers, Television, and New Media like Real People*; Cambridge University Press: Cambridge, UK, 1996; Volume 10, pp. 19–36.
36. Nass, C.; Moon, Y. Machines and Mindlessness: Social Responses to Computers. *J. Soc. Issues* **2000**, *56*, 81–103. [CrossRef]
37. Ehret, J.; Bönsch, A.; Aspöck, L.; Röhr, C.T.; Baumann, S.; Grice, M.; Fels, J.; Kuhlen, T.W. Do Prosody and Embodiment Influence the Perceived Naturalness of Conversational Agents' Speech? *ACM Trans. Appl. Percept.* **2021**, *18*, 1–15. [CrossRef]
38. Nussbaum, C.; Frühholz, S.; Schweinberger, S.R. Understanding Voice Naturalness. *Trends Cogn. Sci.* **2025**, *29*, 467–480. [CrossRef] [PubMed]
39. Diel, A.; Lewis, M. Deviation from Typical Organic Voices Best Explains a Vocal Uncanny Valley. *Comput. Hum. Behav. Rep.* **2024**, *14*, 100430. [CrossRef]
40. Ross, A.; Corley, M.; Lai, C. Is There an Uncanny Valley for Speech?: Investigating Listeners' Evaluations of Realistic Synthesised Voices. In *Proceedings of the Speech Prosody 2024*; International Speech Communication Association (ISCA): Seoul, Republic of Korea, 2024; pp. 1115–1119.
41. Grayson, K.; Martinec, R. Consumer Perceptions of Iconicity and Indexicality and Their Influence on Assessments of Authentic Market Offerings. *J. Consum. Res.* **2004**, *31*, 296–312. [CrossRef]
42. Koles, B.; Audrezet, A.; Moulard, J.G.; Ameen, N.; McKenna, B. The Authentic Virtual Influencer: Authenticity Manifestations in the Metaverse. *J. Bus. Res.* **2024**, *170*, 114325. [CrossRef]
43. Alimardani, M.; de Roode, R.; Vaitonyte, J.; Louwerse, M.M. Effect of a Virtual Agent's Appearance and Voice on Uncanny Valley and Trust in Human-Agent Collaboration. In *Proceedings of the 24th ACM International Conference on Intelligent Virtual Agents*; Association for Computing Machinery: New York, NY, USA, 2024; pp. 1–7.
44. Nowak, K.L.; Biocca, F. The Effect of the Agency and Anthropomorphism on Users' Sense of Telepresence, Copresence, and Social Presence in Virtual Environments. *Presence Teleoperators Virtual Environ.* **2003**, *12*, 481–494. [CrossRef]
45. Mende, M.; Scott, M.L.; Van Doorn, J.; Grewal, D.; Shanks, I. Service Robots Rising: How Humanoid Robots Influence Service Experiences and Elicit Compensatory Consumer Responses. *J. Mark. Res.* **2019**, *56*, 535–556. [CrossRef]
46. Mori, M.; MacDorman, K.F.; Kageki, N. The Uncanny Valley. *IEEE Robot. Autom. Mag.* **2012**, *19*, 98–100. [CrossRef]
47. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2010; pp. 1459–1462.
48. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2015**, *7*, 190–202. [CrossRef]
49. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [CrossRef] [PubMed]
50. Talebi, H.; Milanfar, P. NIMA: Neural Image Assessment. *IEEE Trans. Image Process.* **2018**, *27*, 3998–4011. [CrossRef] [PubMed]
51. Cameron, A.C.; Trivedi, P.K. *Regression Analysis of Count Data*; Cambridge University Press: Cambridge, UK, 2013.
52. Hilbe, J.M. *Negative Binomial Regression*; Cambridge University Press: Cambridge, UK, 2011.
53. Yao, K.; He, K.; Zhang, C. Leveraging Credamo for Efficient Data Collection: Functional Introductions and Cross-Field Applications in Academic and Commercial Settings. *J. Appl. Bus. Behav. Sci.* **2025**, *1*, 232–255. [CrossRef]
54. Park, C.W.; Mothersbaugh, D.L.; Feick, L. Consumer Knowledge Assessment. *J. Consum. Res.* **1994**, *21*, 71. [CrossRef] [PubMed]
55. Miller, E.J.; Steward, B.A.; Witkower, Z.; Sutherland, C.A.M.; Krumhuber, E.G.; Dawel, A. AI Hyperrealism: Why AI Faces Are Perceived as More Real Than Human Ones. *Psychol. Sci.* **2023**, *34*, 1390–1403. [CrossRef] [PubMed]
56. Holz, N.; Larrouy-Maestri, P.; Poeppel, D. The Paradoxical Role of Emotional Intensity in the Perception of Vocal Affect. *Sci. Rep.* **2021**, *11*, 9663. [CrossRef] [PubMed]
57. Moulard, J.G.; Garrity, C.P.; Rice, D.H. What Makes a Human Brand Authentic? Identifying the Antecedents of Celebrity Authenticity. *Psychol. Mark.* **2015**, *32*, 173–186. [CrossRef]

58. Sussman, S.W.; Siegal, W.S. Informational Influence in Organizations: An Integrated Approach to Knowledge Adoption. *Inf. Syst. Res.* **2003**, *14*, 47–65. [CrossRef]
59. Cox, D.; Cox, A.D. Beyond First Impressions: The Effects of Repeated Exposure on Consumer Liking of Visually Complex and Simple Product Designs. *J. Acad. Mark. Sci.* **2002**, *30*, 119–130. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.