



Article

How Visual Framing Strategies Shape Consumer Engagement and Sales in Short-Video Commerce

Xue Pan, Xin Xia and Lei Hou *

School of Management Science and Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China; x.pan@nuist.edu.cn (X.P.); 202312710006@nuist.edu.cn (X.X.)

* Correspondence: l.hou@nuist.edu.cn

Abstract

Short videos have become a dominant format in digital commerce, enabling brands to engage consumers and drive purchases through dynamic and visually rich content. This highlights the need for a more nuanced understanding of visual framing strategies, that is, what elements are shown and how they are presented. Drawing on Cognitive Load Theory, this study explores the impact of visual compositional framing strategies and their dynamics on consumer engagement and sales. Applying a CNN-based deep learning model, 249,043 images (video frames) extracted from 3426 book-related short sales videos on Douyin are classified into one of three categories: functional, contextual, or social, according to the visual composition of the frame. Further econometric modeling reveals distinct effects of such framing categories: functional framing is positively associated with both engagement and sales, contextual framing relates to higher sales only, while social framing relates positively to engagement but negatively to sales. From a dynamic perspective, frequent transitions between framing types within a short video increase visual complexity, which reduces both engagement and sales and moderates the effects of specific framing strategies. These findings advance theoretical understanding of visual framing in dynamic media environments and offer practical insights for designing more effective short video content.

Keywords: short video commerce; visual framing; cognitive load theory; deep learning; consumer engagement; product sales

1. Introduction

Short videos have emerged as a dominant format for digital content consumption due to their concise, engaging, and often entertaining nature. Capitalizing on this trend, many brands and content creators have integrated short videos into their marketing strategies by leveraging platforms such as TikTok, Douyin, and YouTube Shorts, to not only promote products, but also drive direct sales. This integration has given rise to short video commerce, which has become a powerful tool for capturing consumer attention in highly dynamic and competitive digital environments [1–5].

Substantial research attention has been attracted to understand what makes short sales videos effective in stimulating consumer engagement and sales, ranging from platform innovation characteristics, source credibility, to the multimodal design features [6–9]. Among these, visual framing, i.e., what elements are shown and how these elements are presented, plays a critical role in shaping consumer engagement and decision-making. Relevant studies often classify visual content into categories from a semantic perspective,



Academic Editor: Rafael Currás-Pérez

Received: 31 March 2026

Revised: 22 May 2026

Accepted: 22 June 2026

Published: 25 June 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

such as informational versus emotional content [10–13], or from a technical perspective, visual perspective, camera distance, shot type, or video format [14]. These studies demonstrate that visual presentation can have distinctive effects in driving consumer engagement. However, while semantic studies explain what meanings or appeals visual content conveys, and technical studies explain how visual content is captured or produced, they provide a limited understanding of how key visual elements are compositionally arranged within the visual frame. These findings underscore the persuasive power of visual presentation in short videos and motivate further investigation into how visual framing is structured and delivered.

Beyond semantic and technical perspectives, this study focuses on the compositional level of visual framing. Compositional framing concerns the relative prominence and arrangement of key visual elements within a visual scene. In short sales videos, such compositional framing can be reflected in whether the visual emphasis is placed primarily on the product itself, the product within a usage context, or the presenter. Accordingly, this study distinguishes three compositional framing strategies: functional framing, contextual framing, and social framing. These strategies are defined by the compositional prominence of visual elements rather than by the semantic appeal or emotional tone of the message. Functional framing refers to a product-dominant composition that emphasizes the visual prominence of the product itself and directs viewers' attention toward product-related information and features (e.g., Figure 1a). Contextual framing refers to a product-in-context composition that emphasizes the product within surrounding usage scenarios or environmental contexts, helping viewers situate the product within broader consumption experiences (e.g., Figure 1b). Social framing refers to a presenter-dominant composition that emphasizes the presence of presenters and interpersonal cues, directing viewers' attention toward social interaction and presenter-related interpretation (e.g., Figure 1c). Together, these three strategies capture the dominant compositional emphases in short sales videos: product prominence, contextual embedding, and presenter prominence. Although such compositional variations are frequently observed in short sales videos, prior studies have rarely systematically distinguished and theorized these framing strategies.

A second critical gap in the literature lies in the tendency to analyze short videos at the aggregate level, thereby overlooking their temporal structure. Unlike static images, short videos are, in fact, inherently dynamic, consisting of multiple visual frames that continuously transition between different compositional framings, such as product-focused, context-focused, or presenter-focused presentations. However, previous studies predominantly treat each short video as a single static unit, by either manipulating the overall presentation of a video [15,16] or manually coding a video's overall attributes [13,17,18]. While these studies offer valuable insights, they largely overlook the sequential organization of visual framing, that is, how different framing compositions are arranged and transition over time within a single short video. Ignoring this internal dynamic complexity may obscure how viewers process short-video content as a continuous stream of changing visual information. In dynamic short-video environments where viewers process rapidly changing visual cues under limited attentional capacity, frequent transitions between framing compositions may disrupt message coherence, redirect viewers' attention, and increase the cognitive effort required to integrate visual information across frames. As a result, the persuasive effectiveness of short sales videos may depend not only on which compositional framings are presented, but also on how these framings are dynamically organized and transitioned throughout the video.



(a)



(b)



(c)

Figure 1. Examples of framing styles in short sales videos: (a) functional framing, (b) contextual framing, and (c) social framing.

Cognitive Load Theory provides a particularly suitable theoretical lens for explaining these issues because it focuses on how information structures consume individuals' limited cognitive resources during information processing. In short-video commerce, viewers need to process rapidly changing visual cues within a very limited time while simultaneously evaluating product information and deciding whether to engage with or purchase the promoted product. Different compositional framings may impose different cognitive demands by directing attention toward product information, contextual interpretation, or presenter-related cues. Likewise, frequent transitions among framing types may further increase the cognitive effort required to integrate visual information across frames. While theories such as the Elaboration Likelihood Model and the Stimulus–Organism–Response

framework are useful for explaining persuasion routes or general stimulus-response processes, they provide less direct explanation of how dynamic visual compositions impose different processing demands under constrained attention. Therefore, Cognitive Load Theory is especially appropriate for theorizing how compositional visual framing and framing complexity affect consumer engagement and sales in short-video commerce. This lens is also useful for explaining why visual strategies that attract interaction may not necessarily facilitate product-related processing and sales conversion.

Tackling these theoretical gaps, the present study introduces a more nuanced and theoretically informed typology of compositional visual framing for short sales videos, grounded in Cognitive Load Theory. Specifically, we distinguish between three distinct visual framing compositions, including functional framing (Figure 1a), contextual framing (Figure 1b), and social framing (Figure 1c), and identify them at the frame-by-frame level using a CNN-based analytical approach. Accordingly, each video is characterized by the aggregated proportions of these framing types, reflecting the extent to which different framings are employed throughout the video. A dynamic perspective is also adopted to quantify the framing complexity of short sales videos, depicting the frequency of transitions across different framing styles within a video. Guided by this framework, the study addresses the following research questions:

RQ1: *How do different compositional visual framing strategies in short sales videos influence consumer engagement and product sales?*

RQ2: *How does framing complexity, reflected by transitions among different framing strategies, influence consumer engagement and product sales?*

RQ3: *How does framing complexity moderate the effects of specific visual framing strategies on consumer engagement and product sales?*

Overall, this study contributes to the literature on short-video commerce by clarifying how compositional visual framing and dynamic framing complexity shape consumer engagement and sales, while also offering scalable analytical tools and practical guidance for designing more effective short sales videos.

2. Literature Review

2.1. Antecedents of Short Video Commerce Effectiveness

Short video commerce, a dynamic tool within the broader landscape of social commerce, refers to the integration of short videos into the marketing and sales processes, leveraging platforms such as Douyin, TikTok, and YouTube Shorts [19,20]. These videos, typically ranging from a few seconds to a few minutes, convey promotional content in a highly condensed format and often feature direct purchase links. For example, TikTok's "little yellow car" shopping icon enables seamless access to product details and check-out, bridging content consumption with transactional behavior in an entertaining and social environment [3]. This unique combination of entertainment and engagement is reshaping the online shopping experience, offering a compelling alternative to traditional e-commerce models. Unlike conventional online retail, short video commerce captures attention through rich media and platform-specific affordances, and can trigger a desire to postpone closure, keeping users engaged [7]. As a result, short video commerce can effectively foster consumer engagement, including likes, comments, and shares [10,13,21], and such engagement can not only drive onsite purchases [11,17,22] but may also stimulate product sales across platforms [23].

Given its substantial value, a growing body of literature has sought to understand what contributes to the effectiveness of short sales videos in driving consumer engagement

and purchase behavior. As summarized in Table 1, existing studies can be grouped into several streams. The first stream focuses on source- and presenter-related factors, such as presenter attractiveness, credibility, expertise, and tie strength, highlighting the importance of who delivers the promotional message [17,21,24]. The second stream examines social, platform, and interaction-related factors, such as social interaction, entertainment, product visibility, and platform-enabled recommendation mechanisms [7,25]. The third stream investigates semantic content features, including informational, emotional, entertaining, storytelling, and relevance-related appeals [10,11,13,18]. More recently, another stream has begun to examine technical and multimodal design features, such as visual perspective, voice-over source, audio fluctuation, visual brightness, contrast, saturation, shot number, image complexity, and video format [8,15,16,26]. Collectively, these studies show that short-video commerce effectiveness is shaped not only by source, platform, and product factors, but also by how promotional content is visually and multimodally presented. However, prior research has paid comparatively limited attention to the compositional organization of visual elements within frames and the dynamic transitions among different visual compositions over time. This limitation motivates the present study's focus on compositional visual framing and framing complexity as key design features influencing the performance of short sales videos.

Table 1. Studies on the short video commerce effectiveness.

Study	Outcome	Antecedents	Theoretical Framework	Data Source	Approach
[10]	Consumer sharing	Information-focused content, emotion-focused content, brand prominence		YouTube	Data analytics; manual coding
[11]	Consumer engagement and purchase intention	Informative, emotional, and entertaining appeals		Survey	Survey
[15]	Travel intention	Visual framing (first- vs. third-person view, activity- vs. site- centric content), verbal framing (rational vs. emotional)	Construal Level Theory, Elaboration Likelihood Model	Experiment	Experiment
[21]	Consumer engagement	Performance expectancy, entertainment, tie strength, sales approach	Uses and Gratifications Theory, Signal Theory	TikTok	Text analytics
[24]	Travel intention	Presenter, destination, video design attractiveness	Stimuli-Organism-Response	Douyin	Grounded Theory
[13]	Consumer engagement	Informational and emotional content, foreign language	Theory of Exchange, Uses and Gratifications Theory	TikTok	Data crawling; manual coding
[17]	Purchase intention; Sales	Video trustworthiness, expertise, attractiveness, authenticity, brand heritage		TikTok	Grounded Theory; manual coding
[7]	Purchase intention	Social interaction, entertainment, product visibility, recommendation	Sociotechnical Theory	Survey	Survey
[27]	Purchase intention and ad conversion	AI vs. human voice over	Cognitive Load Theory	TikTok, experiment	Data analytics; experiment

Table 1. Cont.

Study	Outcome	Antecedents	Theoretical Framework	Data Source	Approach
[18]	Consumer engagement	Content matching, information relevance, storytelling, and emotionality		Weibo	Data analytics; manual coding
[8]	Consumer engagement	Textual (sentiment), audio (fluctuation, pitch, diversity), visual (brightness, contrast, saturation) signals	Signaling Theory	Douyin, Chanmama	Data analytics; fsQCA
[26]	Consumer engagement	Visual (number of shots, pixel-level image complexity, vertical versus horizontal formats), audio (speech rate, average spectral centroid)	Signaling Theory	TikTok	Data analytics;
This study	Consumer engagement, product sales	Visual framing (functional, contextual, and social framing), framing complexity	Cognitive Load Theory	Douyin, Chanmama	Data analytics, deep learning

2.2. Visual Framing of Product Promotion Content

Framing is a communication process through which message creators emphasize certain aspects of reality to shape how audiences perceive and interpret information. According to Entman [28], framing involves making selected elements of a message more salient “in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation.” By shaping what is included, excluded, or emphasized, framing functions as a powerful tool for guiding attention, constructing meaning, and ultimately shaping behavioral outcomes. While originally developed in the context of political communication and journalism, framing theory has since been widely adopted across disciplines to understand how media messages influence cognition and decision-making. In e-commerce and social commerce, product-related information is increasingly conveyed through digital media in the forms of either text [29,30], images [31–33], or videos [13,26]. As such, the framing of such content, e.g., how product information is structured and presented, plays a pivotal role in capturing attention and stimulating consumer action, especially for the visual content [34,35].

Research on visual framing has examined a range of technical presentation features that shape how audiences process promotional content. In static images, factors such as camera distance and angle can influence their marketing effectiveness. For example, close-up product promotion images are found to be more effective for utilitarian products in inducing rational appeals, while long-shot images are better at inducing emotional appeals for hedonic products [36]. The sense of motion and dynamism can also be implied in static visuals, which can enhance viewers’ mental imagery and further their engagement intention [37,38]. In short videos, these visual effects are further shaped by sequencing, camera movement, and editing pace, which guide the viewer’s attention and processing over time. Gan et al. [15] demonstrated the important role of visual perspectives, indicating that a first-person view in travel videos largely facilitates imagery fluency (i.e., the ease with which viewers can mentally visualize or simulate the depicted scene) and promotes travel intention among viewers. Recent work also identifies video orientation (e.g., vertical versus

horizontal formats) as a technical framing feature that can influence viewer engagement and interaction patterns [26].

Beyond technical presentation, visual framing can also be defined in terms of the semantics that the visual elements convey. For example, the visuals can be classified as product-focused versus experience-focused [14], activity-centric versus site-centric [15], or concrete versus abstract product [39], each demonstrating very different effects on consumer engagement and behavior. Among these, a common approach in the literature is to distinguish between informational versus emotional visual content, for either static images [40] or videos [10,13,41]. Informational content normally refers to visuals that provide factual details and attributes about the promoted product/brand, which tend to support rational processing and drive purchase behavior, especially for tangible, utilitarian, and low-involvement products. On the other hand, emotional content refers to those designed to evoke emotional responses, such as warmth, love, pride, and joy, and can boost consumer engagement, especially in emotionally rich or hedonic consumption contexts. Given the semantic nature of the informational and emotional visuals, most relevant studies rely on manual coding approaches to determine whether content is informational or emotional, rather than directly examining their explicit visual elements. Recent research has proposed more specific operationalization methods for promotional images. For instance, Gu et al. [12] classify images that clearly depict the product or brand as informational, while those featuring human faces with positive expressions as emotional. Chen et al. [42] adopt a similar approach, identifying on-image text as informational content and human faces as emotional content.

Notably, these studies evaluate informational and emotional cues separately, either computationally or manually. As a result, a single stimulus can score high on both dimensions if it contains features that convey factual information and features that evoke emotion. For instance, a presentation of a person with exciting facial expressions using the target product can convey strong informational appeals and, at the same time, strong emotional appeals. Accordingly, the previous classifications on visual framings cannot capture the nuanced differences in visual compositions, such as whether the product is presented in isolation or together with a presenter. Such visual compositions may fundamentally shape how audiences allocate attention, process information, and respond to the persuasive message. Therefore, it is of both theoretical and practical significance to examine how compositional styles, being independent from the semantic or technical feature, influence consumer responses within the dynamic and multimodal environment of promotional videos.

2.3. Cognitive Load Theory

Cognitive Load Theory posits that an individual's cognitive resources are limited, and their ability to process and learn information is shaped by the amount and type of cognitive effort required to engage with a given task [43]. The theory identifies three distinct types of cognitive load: intrinsic, extraneous, and germane. Intrinsic load refers to the inherent complexity of the material being processed, often determined by the number of elements and their interactivity. Extraneous load, by contrast, stems from the inclusion of peripheral or non-essential elements that may distract attention from core information, often interfering with efficient processing and reducing comprehension. Germane load refers to the cognitive effort directed toward schema construction and knowledge integration, and is the beneficial load that supports deep understanding. Together, intrinsic load and extraneous load have negative influences on the efficiency and effectiveness of information processing, while germane load provides a beneficial influence.

The theory has been extensively applied to understand user behaviors in various digital and mediated environments. For example, Van der Land et al. [44] found that increased visual complexity in 3D environments raises extraneous load and hinders comprehension, emphasizing the role of presentation design in shaping user understanding. In e-commerce contexts, research shows that interacting with content in a foreign language significantly increases both intrinsic and extraneous load, resulting in cognitive fatigue and reduced satisfaction [45]. In online review settings, Wang and Li [46] demonstrate that reviews with higher content density (intrinsic load) and poorly formatted presentation (extraneous load) reduce user engagement, while well-matched, relevant content increases germane load and engagement. Likewise, Wen et al. [47] show that under high cognitive load, users are less capable of recognizing persuasive cues, highlighting the theory's relevance to advertising and persuasive communication. In digital environments saturated with information, excessive intrinsic or extraneous load has been found to trigger cognitive overload, leading to anxiety, fatigue, and even avoidance behaviors [48].

For short video commerce, Cognitive Load Theory also offers a valuable lens to understand how viewers interact with highly dynamic, visually rich, and often fast-paced content. In practice, potential consumers need to process short sales videos with their limited cognitive resources to decide whether to engage with and purchase the product being promoted. As such, the intrinsic, extraneous, and germane load of consumers when processing short sales videos may shape their behavioral decisions. For example, as shown in Wang et al. [16], AI-generated voice-overs in short video ads increase extraneous load and reduce purchase intention compared to human voice-overs. However, existing applications of Cognitive Load Theory to short video commerce remain limited, often focusing only on isolated design features (e.g., narration, language) rather than the overall visual structure and presentation style. In particular, different compositional framings may guide viewers' limited cognitive resources toward different information targets, such as product features, usage contexts, or presenter-related cues, thereby imposing distinct processing demands. In addition, frequent transitions among framing types may increase the effort required to integrate visual information across frames, making framing complexity an important source of cognitive load in dynamic video environments.

2.4. Visual Framing and Its Cognitive Load

As highlighted earlier, prior research often distinguishes different semantic types of the visual content [12,13,42] or compares different technical types of how the visuals are captured [26,36]. However, it rarely distinguishes between specific presentation compositions, i.e., which elements to feature to compose the visual frame. Such compositional framing closely correlates to viewers' attention allocation, and thus may impose distinct cognitive demands and shape viewer responses in different ways. Through systematic empirical observations of a large corpus of short sales videos, three dominant and recurrent patterns of visual composition consistently emerged: functional framing (featuring the product alone), contextual framing (embedding the product in a usage environment), and social framing (featuring a presenter speaking without showing the product). By aligning these visual compositions with specific cognitive load categories, this framework offers a theory-driven lens to explain variation in consumer engagement and purchase behavior in short video commerce. The hypothesized effects of these framing types are shown in Figure 2.

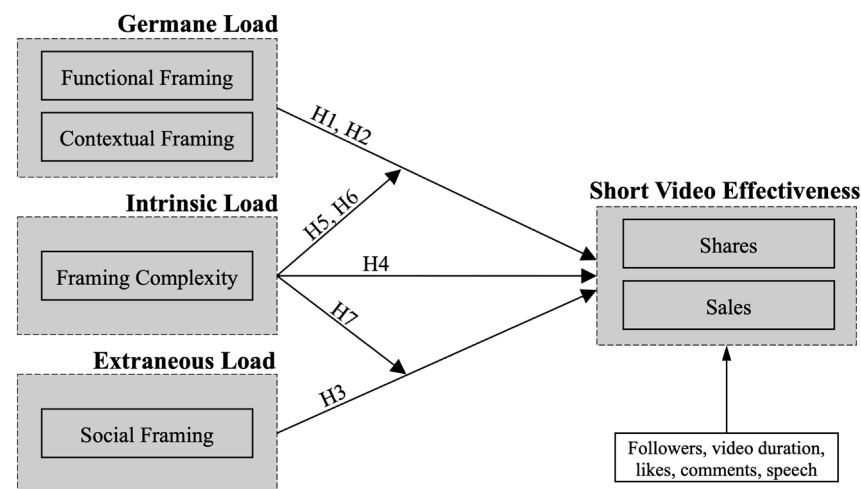


Figure 2. Research framework.

2.4.1. Functional Framing and Germane Load

Functional framing in a short sales video refers to visual presentations that focus directly on the product itself, i.e., its features, attributes, or functionalities. Such presentations are task-relevant and align with the viewer's goal of quickly understanding what the product is and how it works. Because functional framing foregrounds the product itself, it reduces ambiguity about the focal object and directs cognitive resources toward product-related information. This supports germane processing by helping viewers organize and evaluate product attributes more efficiently. The functional framing visuals, from a presentation composition perspective, are different from the widely studied “informational content” in the informational-emotional typology. Rather than broadly including any fact-oriented or rationally appealing content, which could also be delivered by verbal narration [10], functional framing, as defined in this study, emphasizes direct visual depictions of the isolated product, excluding visual contextual or social elements that might divert viewer attention from core product information.

From a Cognitive Load Theory perspective, functional framing primarily contributes to germane load, as it directs cognitive effort toward understanding the product in a structured and task-relevant way. By providing clear and focused information, it supports the viewer's goal of evaluating the product, facilitating schema construction without adding unnecessary complexity [43]. This alignment between content and viewer intention promotes deeper processing. Empirical evidence consistently highlights the persuasive value of such product-focused content. It has been evidenced that when viewers perceive higher informativeness in TikTok ads, they are more likely to engage and purchase the product [11]. As such, informational cues, which could also be delivered in functional framing visuals, have been shown to increase consumer engagement, such as likes, shares, and comments [13,30], as well as drive sales [41,42]. To be more specific, for the functional framing conceptualization, explicitly providing product visuals has also been proven beneficial. Visually detailed product presentations on e-commerce websites enhance consumers' perception of product quality [49]. On social media, firm-generated images featuring the brand name or the product tend to generate stronger consumer engagement [12]. Likewise, short videos in which products are clearly visible are more likely to elicit higher purchase intentions [7]. Collectively, these findings underscore the importance of functional framing in boosting promotional effectiveness. Integrating with the Cognitive Load Theory, functional framing may thus promote consumers' germane load, and further enhance the effectiveness of short sales videos, leading to the following hypothesis:

H1: *Functional framing in short sales videos positively influences (a) consumer engagement and (b) product sales.*

2.4.2. Contextual Framing and Germane Load

Contextual framing in a short sales video refers to visual presentations that highlight a product within a real-life usage scenario or lifestyle setting, allowing viewers to see how the product fits into everyday routines. This approach goes beyond merely displaying the product itself, but instead embeds the product in scenes, such as featuring an individual using the product. Previous informational-emotional classifications might treat such visuals as either informational or emotional, since they could potentially convey both product-related facts and emotionally-rich facial expressions (e.g., [12,40]). However, it is totally different from presenting the product alone or highlighting the facial expressions of individuals alone, from the perspective of visual compositions. As a consequence, contextual framing differs from prior conceptualizations of either informational or emotional content, bridging product information with experiential relevance and offering a stronger link between the product and its intended use. By embedding the product within a usage context, contextual framing connects product information with concrete consumption scenarios. This can facilitate mental simulation and schema construction, thereby supporting germane cognitive processing.

From a cognitive load perspective, contextual framing contributes to germane load by embedding the product in familiar or aspirational usage settings that aid schema construction and decision relevance [43]. Rather than increasing complexity, contextual framing enables viewers to engage in meaningful simulation of how the product might integrate into their own lives. According to prior literature, contextual, immersive, and demonstrative-style product presentations are able to improve imagery fluency and trigger vivid mental simulations, thereby further enhancing purchase intentions [32,50]. Factors related to the current contextual framing have also been explored in tourism and hospitality research. In tourism short video marketing, Gan et al. [15] highlighted the superior effectiveness of activity-centric visuals in promoting travel intention, which feature potential activities at a destination (the focal product) rather than showing the destination alone. Similarly, in the restaurant review context, Luo et al. [51] operationalized the concept of visual breadth (the diversity of the presented visual elements) as the number of distinct content dimensions (environment, food, menu, people) covered by review photos. As such, photos with higher visual breadth place the focal product, i.e., food, within a rich context of environment and consumer dining, thereby enhancing perceived review helpfulness. In the hotel review context, Park and Lee [52] demonstrated that experience-focused consumer photos, those depicting the consumption experience rather than static facilities, elicited significantly stronger visit intention. Building on this empirical evidence and the Cognitive Load Theory, we believe the congruence between product and personal relevance supports deeper processing and improves message effectiveness, thus potentially fostering favorable behavioral outcomes. Accordingly, we propose the following hypothesis:

H2: *Contextual framing in short sales videos positively influences (a) consumer engagement and (b) product sales.*

2.4.3. Social Framing and Extraneous Load

Social framing in a short sales video refers to visual presentations that focus on presenters, such as an influencer, salesperson, or narrator, without showing the product itself. It is conceptually distinct from the widely studied emotional content in previous literature, primarily in two ways. First, while emotional content delivers information that aims to evoke emotional responses, social framing characterizes mainly the visual composition

of human presence. Although some studies operationalized emotional content as images that feature a human face (e.g., [12]), their emphasis is on the emotional expressions (e.g., happiness) conveyed by those faces, rather than on the human presence as a compositional element. Second, emotional content in prior research can be delivered through either verbal or visual means, as long as the content can trigger emotional responses [13,41], whereas social framing in this study is defined specifically by its visual delivery, regardless of whether the presenter's verbal message is emotional or informational.

Such social framing visuals are common in influencer marketing, where the presenter's appearance, persona, and delivery style are central to the viewer experience. However, despite the general effectiveness of including an influencer in visual marketing, it is still critical to distinguish between highlighting the influencer (social framing) solely or emphasizing the product/brand as well (contextual framing). When the presenter becomes the dominant visual element while the product is absent or less salient, viewers may need to integrate presenter-related cues with product information conveyed elsewhere. This can increase split attention and extraneous processing.

Prior research shows that focusing on a presenter can overshadow the brand or product being promoted, producing an eclipsing effect for viewers who are not strongly attached to the presenter [53]. Similarly, Tellis et al. [10] found that video ads featuring a narrator tend to elicit lower consumer engagement, even when the narrator delivers informational product arguments. To counteract this, recent studies suggest making the product visually salient in influencer marketing images and videos, for example, by increasing its brightness and color contrast [54] or positioning it in more attention-grabbing areas of the frame [5], rather than relying solely on the influencer's presence.

From a Cognitive Load Theory perspective, such effects can be explained through the split-attention principle. When viewers are required to match and integrate information across separated sources (e.g., a presenter's verbal description and a product absent from the visuals), their extraneous cognitive load increases significantly, leading to inefficient information processing [27,55]. In addition, poorly aligned information design that fails to direct attention toward task-relevant content may also impose unnecessary load [56]. This aligns with the idea that peripheral social cues, such as expressive presenters in short sales videos, may distract viewers and consume cognitive resources that would otherwise be allocated to processing product information. As such, social framing, by emphasizing solely the presenter rather than the product, may increase extraneous load and reduce the overall effectiveness of the message. Accordingly, the following hypothesis is proposed.

H3: *Social framing in short sales videos negatively influences (a) consumer engagement and (b) product sales.*

2.4.4. Framing Complexity and Intrinsic Load

Unlike static images or texts, short sales videos are inherently dynamic, often composed of multiple shots that shift across different visual compositions, for example, transitioning from functional framing to contextual framing or social framing. From this compositional perspective, such variation reflects framing complexity, defined here as the frequency of transitions across distinct framing types within a single video. These transitions alter how visual elements are arranged and emphasized, thereby changing where viewers' attention is directed over time. Consequently, higher framing complexity indicates greater visual compositional variability, which can increase the cognitive effort required for viewers to maintain attention and integrate information across shifting visual contexts. Frequent transitions among framing types increase the number and interactivity of visual elements that viewers must process over time. This raises integration demands across frames and may increase intrinsic cognitive load. However, prior research has mainly

focused on the overall attributes of short sales videos (e.g., [16,24,26]), overlooking the potential effects of the inherent and dynamic complexity of videos.

According to Cognitive Load Theory, this form of complexity corresponds to intrinsic cognitive load, which stems from the structure and difficulty of the content itself [46]. Prior studies show that when information contains multiple, interacting elements, such as varied product attributes in reviews or dense visual layouts, users experience higher intrinsic load, leading to slower processing and lower engagement [57,58]. In the context of short sales videos, frequent shifts between framing styles may increase the mental effort required for information integration and disrupt perceptual fluency. As the viewer must constantly adjust their focus and interpret new types of visual cues, the processing burden rises, potentially diminishing the effectiveness of the message. Therefore, we have the following hypothesis.

H4: *Framing complexity of short sales videos negatively influences (a) consumer engagement and (b) product sales.*

Cognitive Load Theory also posits that when intrinsic load is high, it becomes especially important to minimize extraneous load and enhance germane load in order to facilitate effective information processing [56]. This suggests that intrinsic load does not operate in isolation, but can interact with other cognitive and contextual factors to shape user experience. For example, Lange and Costley [59] found that as intrinsic load increases, the positive impact of self-regulated effort on germane load also strengthens. In the domain of short video marketing, Wahid et al. [13] show that the use of foreign or code-switched language in TikTok videos negatively moderates the effect of content characteristics on consumer engagement. From a cognitive load perspective, such linguistic complexity reflects elevated intrinsic load, which can disrupt the viewer's ability to process otherwise effective message cues. Extending this logic to visual framing, we argue that when short sales videos are high in framing complexity (intrinsic load), viewers' cognitive resources may become strained, thereby diminishing the positive effects of functional and contextual framing (germane load), while intensifying the negative effects of social framing (extraneous load). Based on this reasoning, we propose the following hypotheses.

H5: *Framing complexity weakens the positive effect of functional framing on (a) engagement and (b) sales.*

H6: *Framing complexity weakens the positive effect of contextual framing on (a) engagement and (b) sales.*

H7: *Framing complexity strengthens the negative effect of social framing on (a) engagement and (b) sales.*

3. Data and Methods

To empirically test the proposed conceptual framework, the study takes books as the focal product and collects short videos that promote and sell books from chanmama.com, China and DouYin, China. A hybrid approach combining manual coding and deep learning is applied to analyze the visual content of short videos to determine their visual framing features. At last, the study develops econometric models to explore the effects of visual framing on the effectiveness of short videos. The overall design of the study is shown in Figure 3.

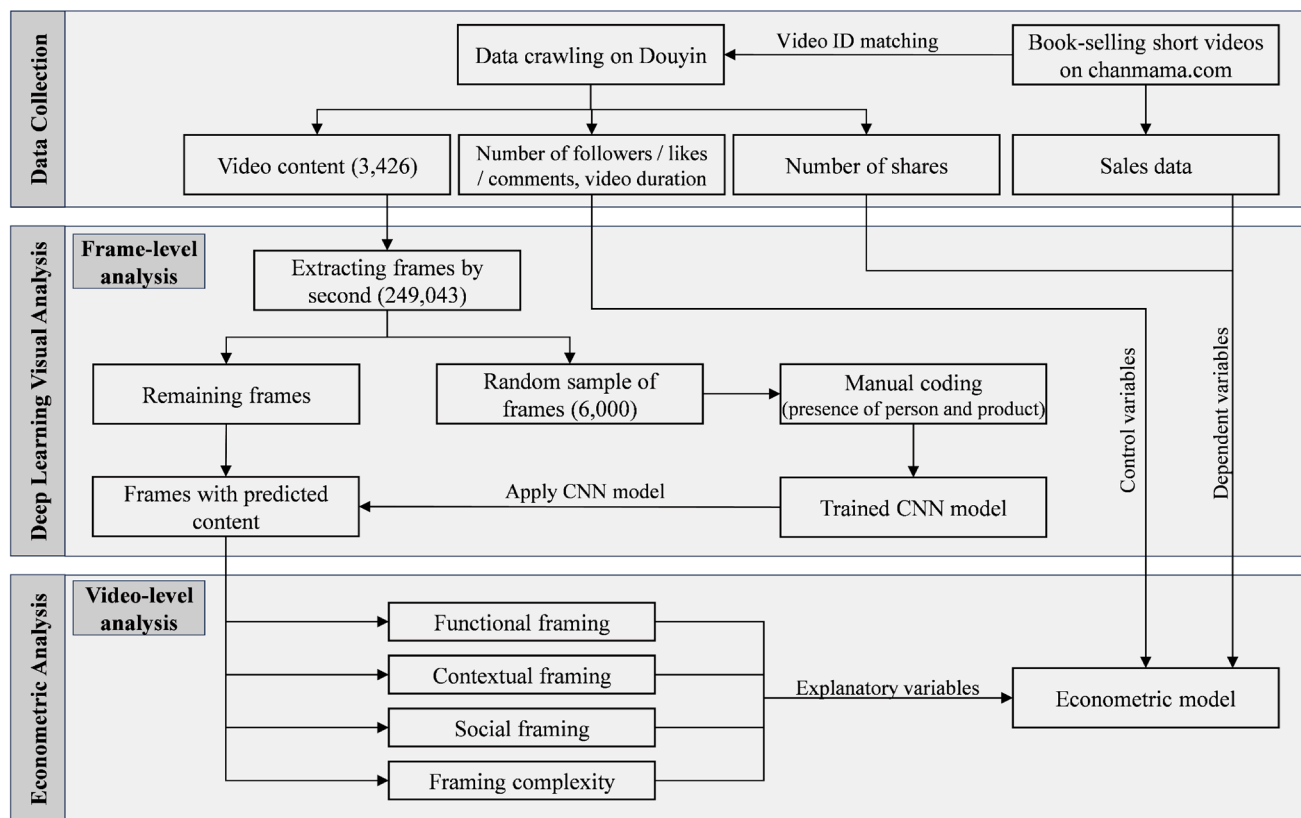


Figure 3. Study design.

3.1. Data Collection

The data was collected from two platforms, namely chanmama.com and DouYin. Chanmama.com is a third-party platform that offers various services of brand content marketing and social commerce, including the monitoring of key metrics of short sales videos on DouYin. The present study focuses on book products because books represent an information- and content-intensive product category in which consumers rely heavily on visual and contextual presentation to infer product value. In addition, short book-promotion videos frequently alternate among product-focused, context-focused, and presenter-focused visual compositions, making them particularly suitable for examining compositional framing and framing transitions. Targeting the book category, we collected data of short sales videos published between January and March 2024 on 10 April 2024. The selected sampling period reflects a relatively stable operational stage of short-video commerce rather than a single short-term promotional campaign, thereby helping reduce the influence of temporary campaign-driven fluctuations. The collected data includes the sales that were directly generated by each short video (purchases made by clicking the sales link within the video) and the hyperlinks to the original short videos on DouYin. Accordingly, a Python 3.0-based web crawler was developed to collect short video data from DouYin by visiting these hyperlinks. Data collected from DouYin includes each video itself and its relevant statistics, such as the number of followers, likes, comments, and shares. After filtering invalid hyperlinks and short videos with incomplete information, the final dataset consists of 3426 short videos promoting books, including their visual content, platform statistics (e.g., followers, comments, shares), and their sales induced directly by the video. Such a dataset enables us to explore the interaction between visual content, consumer engagement, and sales.

3.2. Identification of Visual Framing via Deep Learning

To enable a nuanced exploration of the dynamic feature of short sales videos, we analyze the videos on a frame-by-frame level. Videos are actually a sequence of static images, i.e., frames. Normally, there would be at least 24 frames per second. Although visual impressions can be formed in as little as 100 ms [60], a particular visual framing in short videos, especially higher-level presentation compositions such as product-, context-, or presenter-dominant framing, typically persists for several seconds before changing, even when micro-level editing transitions occur within the sequence. Sampling at the millisecond or sub-second level would therefore capture largely redundant information while exponentially increasing the data volume, resulting in computational infeasibility for large-scale analysis [61]. To balance representativeness and feasibility, we therefore extract one frame per second for each video. For example, for a short video with a duration of 15 s, we use 15 frames (images) to represent the content. Accordingly, the present study focuses on higher-level compositional transitions rather than frame-by-frame cinematographic variations. In this way, 249,043 frames were extracted from the collected short videos.

Each frame of the short video contains rich visual content. Previous computational studies have relied on the specific elements that appear in the visuals to determine the visual framing. For example, visuals containing a specific product, the name of a brand, and texts can be regarded as informational content, while those featuring human faces can be regarded as emotional content [12,42]. While visual framing may also involve finer-grained features such as emotional expression, cinematographic style, aesthetic presentation, or text overlays, the present study specifically focuses on the compositional prominence of key visual elements within a frame. Accordingly, we classify a given frame according to the presence of two commonly observed visual elements, namely person (the presenter) and book (the product). To be more specific, a frame is classified as functional if it presents only a book, as contextual if it presents both a book and a person, and as social if it presents only a person. These categories are intended to operationalize whether viewers' attention is directed primarily toward the product itself, the product within a usage context, or the presenter. This simple yet composition-oriented rule reflects how attention within a frame is visually distributed across social and functional elements. Accordingly, the purpose of the present study is not to capture all possible visual rhetorical features, but to operationalize a scalable measure of high-level compositional prominence in short sales videos. As such, a key step for identifying the presentation composition of a frame is to determine whether a person/book is present.

While we aim to build a deep learning method to identify the presentation type of each frame, such a method requires training data with clear labels. Accordingly, we randomly sampled 6000 frames for manual coding. The coding was done by a group of 40 undergraduate students, each labeled 450 images. Each coder was required to label the images from two perspectives, i.e., whether the image contains a person and whether the image contains a book. Before coding, a brief training was offered to the coders with example images. Each image was coded by three independent coders. To test the inter-coder reliability of the data annotation, we apply Krippendorff's alpha for nominal data, based on the coding results from three independent coders per image. This coefficient quantifies the degree of agreement among coders beyond chance, providing a robust indicator of coding consistency. The resulting Krippendorff's alpha values indicate satisfactory reliability, with 0.920 for person annotation and 0.793 for book annotation, reflecting strong and acceptable levels of agreement, respectively. For images where all three coders provided consistent annotations, the final label was determined directly from the unanimous results. For images with inconsistent annotations, a group-coding procedure was implemented to ensure transparency and accuracy. In this stage, all images with disagreements were

re-examined collectively by the coding team, and the final label was decided through discussion until consensus was reached. Such manual coding finally yields a dataset of 6000 images with labels of whether a book/person is present.

The study further develops a deep learning model based on the Convolutional Neural Network (CNN), as shown in Figure 4. The model takes an input image with a 72×128 resolution in grayscale format. The model integrates 4 convolutional layers and 4 pooling layers. The initial convolutional layers are equipped with 64 3×3 filters and use the ReLU activation function, followed by a max-pooling layer to effectively reduce spatial dimensions. This “convolution-pooling” alternating pattern runs throughout the network, and as the layers deepen, the number of filters gradually increases to 128, 256, and finally 512, allowing the model to progressively extract and refine hierarchical features from the input image. After these convolutional and pooling layers, the resulting feature maps are flattened and passed into a fully connected layer with 1024 neurons and a ReLU activation function. To effectively mitigate overfitting, a Dropout layer is added after this layer. Finally, the output layer consists of two neurons, both with a sigmoid activation function, responsible for generating the classification results, i.e., whether or not a book is present and whether or not a person is present, respectively.

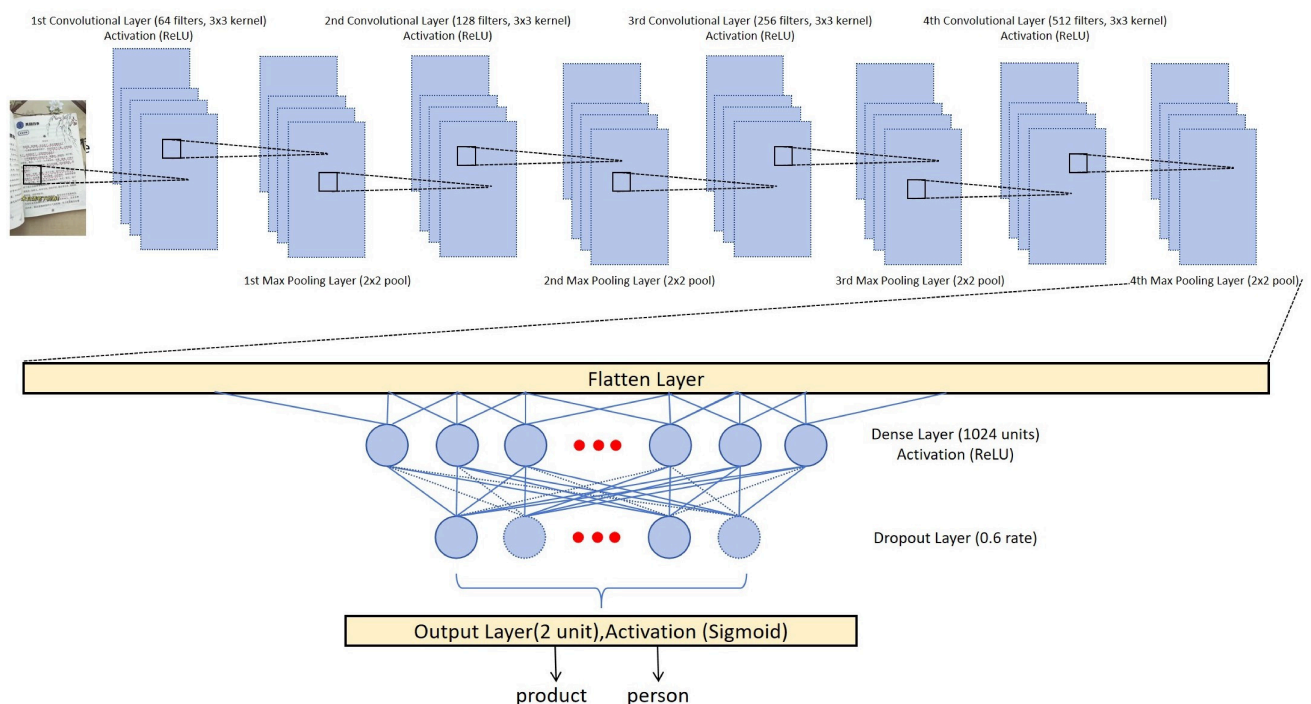


Figure 4. CNN-based model for identifying the visual framing of an image.

To train and test the developed model, the 6000 images with manual labels were randomly divided into a training set (60%), a validation set (20%), and a testing set (20%). The model was trained using the Adam optimizer with a learning rate of 0.001 and binary cross-entropy loss over 30 epochs with a batch size of 32. Over 10 independent data partitions and model testing, the model achieves an average precision of 0.719 and a recall of 0.830 for identifying the presence of a book, and a precision of 0.883 and a recall of 0.866 for identifying the presence of a person, demonstrating the model’s effectiveness. To ensure transparency in model selection, we further compared different architectures, including logistic regression, random forest, SVM, and CNNs with two to five convolutional layers (Table 2). The results indicate that the 4-layer CNN achieves the best balance of precision and recall across both tasks, outperforming shallower CNNs and avoiding the performance

degradation observed in deeper structures. Accordingly, we further apply the trained 4-layer CNN model to all unlabeled frames to predict the presence of a person and a book. With such predictions, we are enabled to determine the presentation type of each frame, to be either functional, contextual, or social.

Table 2. Comparison of classification performance across models over 10 independent runs.

Model	Presence of a Person		Presence of a Book	
	Precision	Recall	Precision	Recall
Logistic Regression	0.687	0.546	0.643	0.720
Random Forest	0.817	0.767	0.790	0.854
SVM	0.821	0.813	0.791	0.871
CNN (2 layer)	0.856	0.824	0.787	0.790
CNN (3 layer)	0.815	0.758	0.778	0.851
CNN (4 layer)	0.883	0.866	0.791	0.830
CNN (5 layer)	0.871	0.822	0.757	0.769

While more advanced deep learning architectures may further improve classification performance, the present study primarily aims to operationalize high-level compositional framing at scale rather than optimize computer vision performance itself. Because the task focuses on detecting broad compositional prominence (i.e., product-, context-, or presenter-dominant compositions), the applied CNN-based model provides a suitable balance between interpretability, scalability, and classification accuracy for the research purpose of the study. Moreover, although classification errors may still exist, such errors are more likely to introduce random measurement noise rather than systematically bias the estimated relationships.

3.3. Variable Operationalization and Model

While the identification of visual framing types was conducted at the frame level, the formal econometric analysis adopts the individual short sales video as the unit of analysis. Accordingly, all variables are operationalized and analyzed at the video level.

In the present study, the focal dependent variables of the present study include shares and product sales. Share ($shares_i$) is used to operationalize an active dissemination-oriented form of consumer engagement in short-video commerce, which is directly collected from DouYin. Sales ($sales_i$) is measured using the midpoint of the platform-provided sales range and is therefore interpreted as a midpoint-based indicator of relative sales performance, rather than an exact transaction count.

The key explanatory variables are the three types of visual framing, i.e., functional framing ($functional_i$), contextual framing ($contextual_i$), and social framing ($social_i$). These are defined at the video level, reflecting aggregated volumes of each framing type within each short sales video. Since each short video has multiple frames, which could be of different types, we use the number of frames that are subject to a particular type as the value to reflect the extent to which the short video presents content of that framing type. For example, if a short video i with 20 frames (20 s duration) contains 10 frames of functional presentation, we have $functional_i = 10$. To reflect the dynamic feature of short videos, the study also considers the framing complexity ($complexity_i$), defined as the number of transitions across different framing types within a short video. As shown in Figure 5, a short video has a sequence of multiple frames. While the first 6 frames could be with a social presentation style, the presentation could become functional starting from the 7th frame, and this can be defined as a transition in framing type. Accordingly, the number of such transitions is able to reflect the complexity of the short videos in terms of the usage of mixed and unstable framing strategies. For example, the short video as illustrated in Figure 5

has $functional_i = 13$ (separated in two video segments), $contextual_i = 6$, $social_i = 6$, and $complexity_i = 3$.

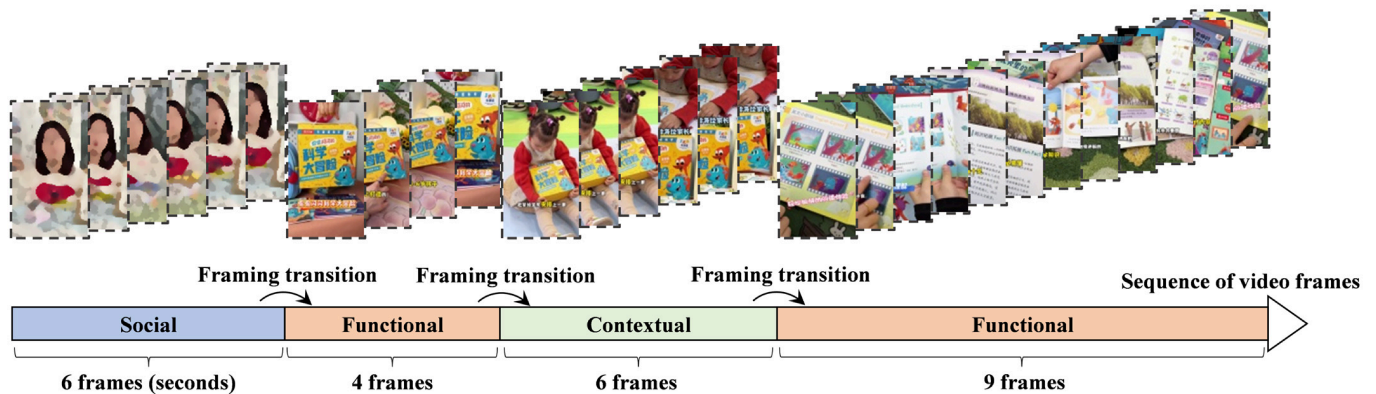


Figure 5. Illustration for framing transitions in short sales videos.

A number of control variables are also considered. On the user level, we control the number of followers of the short video creator ($followers_u$). On the video level, the posting date ($days_i$), i.e., the number of days since the video was posted, number of likes ($likes_i$), and number of comments ($comments_i$) are considered. In addition, most videos have verbal narratives to deliver the product information. Accordingly, we extract text from the audio. On the one hand, we control the spoken words ($words_i$), which is the number of words in the extracted text. On the other hand, SnowNLP is applied to quantify the sentiment polarity of the text ($sentiment_i$), which is also considered a control variable. Finally, we also take the time duration of each short sales video i in seconds ($length_i$) as a control variable. A summary of variable operationalization and brief statistics can be found in Table 3.

In the econometric analysis, we apply z-score normalization to all the independent and control variables. For the variables with highly skewed distributions and extreme values, including functional framing, contextual framing, social framing, framing complexity, number of likes, number of followers, number of comments, days since posting, spoken words, and video length, logarithms are taken before normalization.

Considering the fact that the number of shares is count numbers, a negative binomial model is applied, while for the product sales, which are continuous values, a generalized linear model is implemented on the logarithmic value of product sales. Accordingly, the following models are applied to test the proposed hypotheses:

$$\ln(shares_i) = \beta_0 + \beta_1 functional_i + \beta_2 contextual_i + \beta_3 social_i + \beta_4 complexity_i + \beta_5 complexity_i \times functional_i + \beta_6 complexity_i \times contextual_i + \beta_7 complexity_i \times social_i + \gamma \cdot Control \quad (1)$$

$$sales_i = \beta_0 + \beta_1 functional_i + \beta_2 contextual_i + \beta_3 social_i + \beta_4 complexity_i + \beta_5 complexity_i \times functional_i + \beta_6 complexity_i \times contextual_i + \beta_7 complexity_i \times social_i + \gamma \cdot Control \quad (2)$$

Table 3. Variable statistics.

Variable	Notation	Definition	Mean	Min	Max
Dependent Variables					
Consumer shares	$shares_i$	Number of shares of short video i	921.51	0	243,946
Product sales	$sales_i$	Sales generated by short video i	1036.94	13	62,500

Table 3. Cont.

Variable	Notation	Definition	Mean	Min	Max
Independent Variables					
Functional framing	$functional_i$	Number of video frames that subject to functional framing type in short video i	24.15	0	1635
Contextual framing	$contextual_i$	Number of video frames that subject to contextual framing type in short video i	12.55	0	677
Social framing	$social_i$	Number of video frames that subject to social framing type in short video i	28.44	0	1527
Framing complexity	$complexity_i$	Number of transitions across different framing types in short video i	10.41	0	193
Control variables					
Number of likes	$likes_i$	Number of likes of short video i	4933	0	1,945,377
Number of comments	$comments_i$	Number of comments to short video i	102.60	0	28,055
Number of followers	$followers_u$	Number of followers of video creator u	553,987	97	37,498,886
Days since posting	$days_i$	Number of days since the short video i was posted to the data collection date	32.13	10	69
Verbal sentiment	$sentiment_i$	Sentiment score of text converted from the verbal narratives of short video i	0.89	0	1
Spoken words	$words_i$	Number of words in the verbal narratives of short video i	332.78	0	6956
Video length	$length_i$	Time duration of short video i in seconds	71.515	5.02	1646.04

4. Results

4.1. Descriptive Analysis of Visual Framing

To provide an overview of the structures of short sales videos in terms of the visual framing, we begin with a descriptive analysis. Figure 6 shows the distribution in boxplots of the number of frames using each of the three identified visual framing strategies, as well as the number of transitions between them, i.e., the framing complexity. Functional framing and social framing appear most frequently across short videos. In particular, social framing presents a wide distribution and a long upper tail, suggesting that some videos rely heavily on presenter-focused shots. In contrast, contextual framing that embeds products within a usage scenario appears less frequently. Overall, this result indicates that while product and presenter shots are common in short sales videos, scenario-based presentations, where a person is using the product (in this case, reading the book), are applied less prominently. To be more specific, among the 3348 short sales videos analyzed, 91.0% include functional framing, 69.2% include contextual framing, and 73.5% include social framing.

The majority of short videos employ multiple framing strategies. Specifically, 50.7% of videos include all three framing types, 32.3% incorporate two of the framing types, and only 17.0% use only a single framing type. Such results highlight the complex and multimodal nature of short sales videos in practice, and underscore the importance of accounting for framing complexity in understanding video effectiveness. In fact, the number of transitions varies largely across different short sales videos, ranging from no transition at all to as many as 193 framing transitions within a single short video. Across all videos analyzed, the number of transitions (framing complexity) has an average value of 10.4 and a median value of 7, suggesting frequent transitions in most short sales videos.

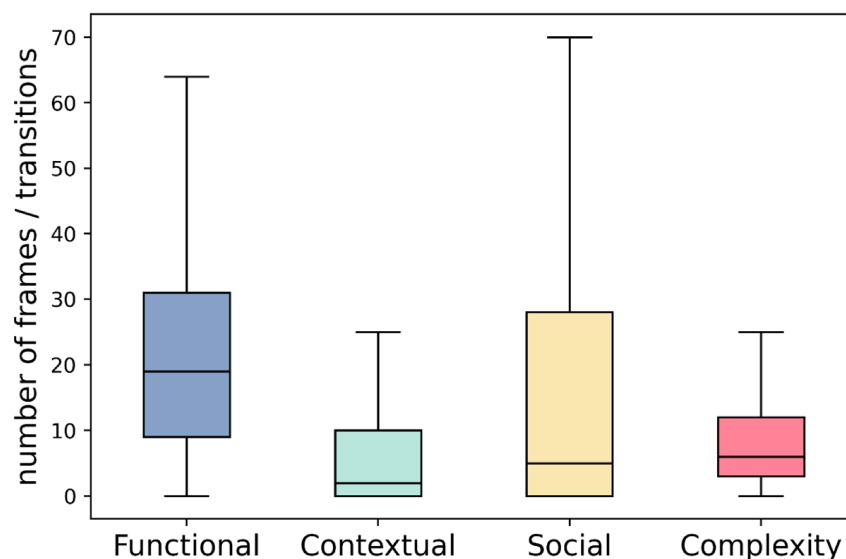


Figure 6. Distributions of the focal variables.

4.2. Main Effects of Visual Framing

To evaluate how different visual framing strategies influence the effectiveness of short sales videos, we apply econometric models as specified in Section 3.3, the results of which are reported in Table 4. Baseline models (Model 1 and Model 4) with only control variables are firstly tested. Among the controls, the number of followers and days since posting exhibit negative effects, which are unexpected. At the bivariate level, the number of shares has positive Pearson correlations with both the number of followers ($\rho = 0.187$, $p < 0.001$) and days since posting ($\rho = 0.252$, $p < 0.001$). The product sales have significant Pearson correlations with days since posting ($\rho = 0.178$, $p < 0.001$), while its correlation with the number of followers is insignificant. However, when the effects of likes and comments are controlled for in the multivariate models, the effects of followers and days since posting become negative, probably due to their strong correlations with likes and comments. Nonetheless, all VIF values remain well below conventional thresholds. It is also worth noting that we consider video length as a control only for baseline models. For other models, video length is excluded because it equals the sum of functional, contextual, and social framings, and including it would create redundancy and multicollinearity.

Table 4. Main effects of visual framing on the effectiveness of short sales videos.

	Consumer Shares (Negative Binomial Model)				Product Sales (OLS Model)			
	Model 1	Model 2	Model 3	VIF	Model 4	Model 5	Model 6	VIF
Functional		0.273 *** (0.021)	0.248 *** (0.024)	2.261		0.251 *** (0.030)	0.215 *** (0.033)	2.260
Contextual		0.018 (0.030)	0.095 ** (0.034)	4.684		0.154 *** (0.041)	0.187 *** (0.047)	4.667
Social		0.301 *** (0.028)	0.320 *** (0.029)	3.503		−0.220 *** (0.039)	−0.202 *** (0.041)	3.461
Framing Complexity		−0.153 *** (0.035)	−0.221 *** (0.037)	5.854		−0.299 *** (0.049)	−0.343 *** (0.053)	5.777

Table 4. Cont.

	Consumer Shares (Negative Binomial Model)				Product Sales (OLS Model)			
	Model 1	Model 2	Model 3	VIF	Model 4	Model 5	Model 6	VIF
Functional ²			−0.012 (0.013)	1.851			−0.048 ** (0.019)	1.849
Contextual ²			−0.066 *** (0.016)	1.554			−0.002 (0.022)	1.562
Social ²			−0.067 *** (0.018)	1.274			−0.025 (0.025)	1.267
Number of likes	2.353 *** (0.032)	2.305 *** (0.031)	2.304 *** (0.031)	3.451	1.587 *** (0.042)	1.520 *** (0.041)	1.514 *** (0.041)	3.552
Number of comments	0.058 (0.030)	0.072 * (0.029)	0.077 ** (0.029)	3.348	0.003 (0.042)	0.052 (0.041)	0.058 (0.041)	3.443
Number of followers	−0.219 *** (0.017)	−0.185 *** (0.017)	−0.178 *** (0.017)	1.183	−0.310 *** (0.025)	−0.273 *** (0.024)	−0.270 *** (0.024)	1.194
Days since posting	−0.042 * (0.017)	−0.018 (0.016)	−0.022 (0.016)	1.099	−0.100 *** (0.024)	−0.106 *** (0.023)	−0.108 *** (0.023)	1.104
Verbal sentiment	0.035 * (0.017)	0.050 ** (0.017)	0.044 ** (0.017)	1.129	0.059 * (0.024)	0.051 * (0.023)	0.048 * (0.023)	1.134
Spoken words	−0.028 (0.027)	−0.129 *** (0.022)	−0.093 *** (0.023)	2.176	0.039 (0.037)	−0.023 (0.031)	0.002 (0.032)	2.162
Video length	0.030 (0.026)				−0.355 *** (0.036)			
Constant	4.240 *** (0.016)	4.216 *** (0.016)	4.358 *** (0.031)		4.846 *** (0.023)	4.846 *** (0.022)	4.921 *** (0.043)	
Observations	3426	3426	3426		3426	3426	3426	
Log Likelihood	−18,099.36	−17,985.47	−17,968.24		−5839.144	−5712.969	−5708.935	
Akaike Inf. Crit.	36,214.720	35,992.940	35,964.480		11,696.290	11,449.940	11,447.870	
Adjusted R ²					0.567	0.598	0.598	
Residual Std. Error					1.332	1.284	1.283	
F Statistic					642.321 ***	509.672 ***	393.253 ***	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$. ² suggest nonlinear relationships.

Overall, the results suggest that the visual framing strategies play very different and distinct roles in driving short sales videos' effectiveness. Functional framing is consistently and significantly associated with higher levels of both consumer engagement ($\beta = 0.273$, $p < 0.001$, Model 2) and product sales ($\beta = 0.251$, $p < 0.001$, Model 5). Such results indicate that visually emphasizing the product itself facilitates goal-relevant processing, i.e., germane load, which further promotes behavioral decisions of both sharing and purchasing. Accordingly, H1a and H1b are fully supported. Contextual framing is revealed to be irrelevant to the consumer shares (Model 2), but significantly correlates with product sales ($\beta = 0.154$, $p < 0.001$, Model 5). As such, only H2b is supported, which suggests that embedding a product in a real-life scenario can effectively drive purchase intention, though not necessarily increase short video virality. Interestingly, social framing has a divergent effect across outcomes. While it positively predicts consumer engagement ($\beta = 0.301$, $p < 0.001$, Model 2), it negatively impacts product sales ($\beta = -0.220$, $p < 0.001$, Model 5). While social framing that highlights a presenter was hypothesized to increase extraneous load that distracts consumer attention, it exhibits an unexpected positive effect

on consumer engagement. This may stem from the parasocial relationships and social identity cues triggered by the engaging presence of the presenter, which encourage viewers to interact with the content. However, these parasocial interactions do not translate into product sales, as they do not provide core product information. These results thus reject H3a but support H3b.

Consistent with our expectations, framing complexity is negatively associated with both engagement ($\beta = -0.153$, $p < 0.001$, Model 2) and sales ($\beta = -0.299$, $p < 0.001$, Model 5). This supports H4a and H4b, indicating that frequent transitions across different framing types may impose greater intrinsic cognitive load, making it more difficult for viewers to process the message effectively.

Beyond the above-discussed linear effects, several nonlinear relationships indicate inverted-U-shaped patterns. Functional framing exhibits a significant negative quadratic effect on sales ($\beta = -0.048$, $p < 0.01$, Model 6), suggesting that while product-focused visuals enhance purchase conversion at moderate levels, excessive reliance reduces such effectiveness. Similar inverted-U-shaped relationships are also found for contextual and social framing on consumer shares. While contextual and social framing are, in general, beneficial for promoting shares, such benefits diminish when they are overly presented, as indicated by the negative quadratic effects of contextual framing ($\beta = -0.066$, $p < 0.001$, Model 3) and social framing ($\beta = -0.067$, $p < 0.001$, Model 3). Together, these nonlinear results highlight that each framing strategy has an optimal range, with overuse creating diminishing or adverse returns.

4.3. Moderation Effects of Framing Complexity

The complexity of visual framing is posited to moderate the effects of different framing styles on the effectiveness of short sales videos. However, the empirical results, as shown in Table 5, only partially support such a proposition.

The results show significant negative interaction between framing complexity and functional framing for both consumer shares ($\beta = -0.030$, $p < 0.05$, Model 7) and product sales ($\beta = 0.088$, $p < 0.001$, Model 8), supporting H5a and H5b. As such, functional framing is more effective when the short sales videos are simply structured, i.e., with fewer transitions across different framing styles. When the video becomes more complex, the persuasive and informational benefits of functional framing appear to weaken. On the other hand, the effects of contextual framing on either consumer shares or product sales are not moderated by framing complexity, rejecting H6a and H6b. While contextual framing has a positive main effect to promote product sales, such an effect is thus persistent regardless of whether the short sales video is complex or not in terms of framing transitions. For social framing, while the complexity does not moderate the effect of social framing on sales, their interaction shows a negative effect for consumer shares ($\beta = -0.070$, $p < 0.001$, Model 7). This implies that social framing is more effective in simpler videos, thereby supporting H7a.

4.4. Robustness Tests

Here, we further test the robustness of the observed effects of different visual framings, mainly from two perspectives, namely the alternative engagement metric and short video length.

In the above analysis, we have considered the number of shares as a key metric for consumer engagement. However, the number of likes and comments, which are taken as controls in the above analysis, are also important consumer engagement indicators in prior empirical studies. As such, here we use the number of likes (Model 9) and the average value of shares and likes (Model 10) to replace the original dependent variable (i.e., number of shares), respectively, and retest the model. We do not consider the number of comments as

an alternative engagement metric, since the comments could be either positive or negative, with potentially different engagement meaning. As indicated by the results in Table 6, the models with alternative engagement metric (Model 9 and Model 10) show consistent effects of the main variables to the original model 2, which takes consumer shares as the focal engagement metric.

Table 5. Moderation effects of framing complexity on framing styles.

	Model 7 (Consumer Shares)		Model 8 (Product Sales)	
	Estimate	Std. Err.	Estimate	Std. Err.
Complexity × Functional	−0.030 *	0.015	−0.088 ***	0.021
Complexity × Contextual	−0.020	0.017	0.042	0.023
Complexity × Social	−0.070 ***	0.018	−0.015	0.026
Functional	0.292 ***	0.022	0.276 ***	0.031
Contextual	0.043	0.03	0.171 ***	0.042
Social	0.313 ***	0.029	−0.183 ***	0.041
Framing complexity	−0.157 ***	0.035	−0.341 ***	0.050
Number of likes	2.312 ***	0.031	1.513 ***	0.041
Number of comments	0.070 *	0.029	0.051	0.041
Number of followers	−0.183 ***	0.017	−0.267 ***	0.024
Days since posting	−0.021	0.016	−0.106 ***	0.023
Verbal sentiment	0.046 **	0.017	0.048 *	0.023
Spoken words	−0.138 ***	0.022	−0.03	0.031
Constant	4.269 ***	0.02	4.827 ***	0.028
Observations	3426		3426	
Log Likelihood	−17,972.180		−5700.452	
Akaike Inf. Crit.	35,972.360		11,430.900	
Adjusted R ²			0.600	
Residual Std. Error			1.280	
F Statistic			396.508 ***	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

The collected short videos have varied lengths, ranging from 5s to 1646s. While we have considered all the videos in the earlier analysis, here we remove the “long videos” and keep those with a length shorter than 120 s ($N = 2947$). Based on such a “very-short” video sample, the variables’ effects on consumer shares (Model 11) and product sales (Model 12) remained consistent with the results of the full sample, indicating robust effects of the different framing types and framing complexity.

Table 6. Robustness test results.

	Alternative Engagement Metric				Very-Short Video Sample (<120 s)			
	Model 9 (Number of Likes)		Model 10 (Shares + Likes)		Model 11 (Consumer Shares)		Model 12 (Product Sales)	
	Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.	Estimate	Std. Err.
Functional	0.397 ***	0.027	0.442 ***	0.027	0.322 ***	0.025	0.365 ***	0.035
Contextual	−0.067	0.038	−0.045	0.038	0.045	0.033	0.268 ***	0.046
Social	0.238 ***	0.036	0.280 ***	0.037	0.357 ***	0.033	−0.107 *	0.045
Framing complexity	−0.090 *	0.045	−0.123 **	0.045	−0.152 ***	0.04	−0.450 ***	0.055
Number of likes					2.266 ***	0.034	1.470 ***	0.044
Number of comments	1.661 ***	0.022	1.683 ***	0.022	0.116 ***	0.032	0.125 **	0.044

Table 6. Cont.

	Alternative Engagement Metric				Very-Short Video Sample (<120 s)			
	Model 9 (Number of Likes)		Model 10 (Shares + Likes)		Model 11 (Consumer Shares)		Model 12 (Product Sales)	
Number of followers	0.025	0.022	−0.005	0.022	−0.189 ***	0.018	−0.263 ***	0.025
Days since posting	0.171 ***	0.021	0.169 ***	0.021	−0.012	0.018	−0.079 **	0.025
Verbal sentiment	−0.297 ***	0.021	−0.300 ***	0.022	0.058 **	0.019	0.057 *	0.027
Spoken words	0.045	0.029	0.04	0.029	−0.130 ***	0.024	−0.019	0.033
Constant	6.802 ***	0.02	6.969 ***	0.02	4.252 ***	0.018	4.891 ***	0.024
Observations	3426		3426		2947		2947	
Log Likelihood	−26,597.340		−27,153.680		−15,378.530		−4886.303	
Akaike Inf. Crit.	53,214.680		54,327.360		30,779.060		9796.606	
Adjusted R ²							0.615	
Residual Std. Error							1.273	
F Statistic							470.705 ***	

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

5. Conclusions and Discussions

5.1. Conclusions

Short sales videos have become an increasingly prominent approach for brands to engage consumers and promote sales due to the visual-rich features. Considering the significance of visual content in short sales videos, the present study examined how visual framing compositions and their complexity influence the effectiveness of short videos in driving consumer engagement and product sales. Drawing on Cognitive Load Theory, the study distinguished three types of framing compositions, including functional, contextual, and social framing, and quantified the complexity of such framings as the frequency of transitions across these three framing styles within a video. The frame-by-frame analysis of 3426 short videos promoting books revealed several key findings as summarized below.

First, with respect to the objective of understanding the differentiated effects of functional, contextual, and social framings, the study finds that functional framing visuals are associated with both higher engagement and increased sales, contextual framing is found to be positively associated with sales, whereas social framing is positively associated with engagement but reduces sales. The divergent effects of social framing are noteworthy and can be interpreted as an engagement–conversion divergence. On the one hand, presenter-centered visuals may provide interpersonal cues, social presence, and affective appeal, thereby encouraging viewers to share or interact with the video. On the other hand, when the presenter becomes the dominant visual element, viewers' attention may shift away from product-related information, making it harder to evaluate the product and form purchase intention. From a Cognitive Load Theory perspective, social framing may therefore enhance interaction-oriented engagement while simultaneously imposing extraneous processing demands that weaken product-focused decision processing. This helps explain why social framing can enhance dissemination-oriented engagement but reduce sales performance. In addition to such general linear effects, inverted U-shaped nonlinear relationships are also observed, indicating that moderate use of each framing is most effective, while overuse leads to diminishing or adverse effects.

Second, in line with the objective of revealing the effects of dynamic framing complexity, we find that higher complexity reduces both engagement and sales directly, yet its moderating role varies by framing type. As expected, complexity weakens the positive effect of functional framing on both consumer engagement and sales. Such moderation is not found for the effects of contextual framing. While social framing is found to promote

consumer engagement, such an effect is also more prominent in simpler videos. These findings highlight that the effectiveness of framing strategies depends not only on their type but also on the structural dynamics of the video.

Collectively, these findings offer a nuanced understanding of how visual framing strategies impose distinct cognitive demands and yield divergent behavioral outcomes of short video commerce.

5.2. Theoretical Contributions

This study makes several theoretical contributions to the literature on short video commerce, visual framing, and the application of Cognitive Load Theory in digital marketing contexts.

First, this study refines the conceptualization of visual framing in product promotion content by introducing a visual compositional perspective. Previous studies have predominantly defined visual framing along semantic dimensions, such as distinguishing informational and emotional content [12,42], or technical dimensions, such as camera angle or distance [26,36]. The present study instead focuses on the compositional arrangement of visual elements within short sales videos, distinguishing whether the visual emphasis lies on the product itself, the product within its contextual environment, or the presenter. This compositional perspective is cognitively relevant because it captures how visual salience and element arrangement shape viewers' attention allocation and information processing. The empirical analysis reveals the divergent effects of these framing types on consumer behavior. As such, the effectiveness of short sales videos depends not only on what is presented, but also on how it is visually composed. By moving beyond broad semantic categories, this framework enables a more precise and cognitively grounded understanding of how visual composition functions as a framing mechanism in digital commerce.

Second, the study applies Cognitive Load Theory to the domain of short video commerce to provide a cognitive explanation for how visual elements shape consumer behavior. While prior applications of the theory in marketing have primarily focused on textual and auditory cues [16,46], little attention has been given to how visual elements contribute to intrinsic, extraneous, or germane load. This omission is particularly critical in the context of short sales videos, where visual elements dominate the user experience. Existing literature has demonstrated the varying effects of different visual elements on consumer behavior [15,26], yet often lacks a cognitive explanation for why and how these effects occur. The findings extend Cognitive Load Theory to dynamic media environments by showing that cognitive load is shaped not only by individual visual elements but also by how visual compositions are organized and transitioned over time. Framing complexity represents a dynamic source of cognitive load because frequent shifts among functional, contextual, and social framing require viewers to repeatedly reallocate attention and integrate changing visual cues across frames. This suggests that, in short-video commerce, cognitive load should be understood not only as a property of isolated information elements, but also as an outcome of their temporal organization. Moreover, the study reveals that engagement and purchase behavior are driven by different types of visual framing, diverging from many previous studies that examine only a single behavior of either engagement (e.g., [13,18]) or purchase (e.g., [7,16]). From a Cognitive Load Theory perspective, this suggests that engagement may be more susceptible to affective or peripheral cues, such as social framing that imposes extraneous load, while purchase behavior depends on deeper and goal-relevant processing as facilitated by functional and contextual framing aligned with germane load. This distinction offers important theoretical insight into the cognitive underpinnings of different consumer behaviors in the context of short video commerce.

Third, this study adopts a dynamic perspective to the analysis of short video content, highlighting the importance of the temporal structure of visual framing. Most prior research focuses on video features at the aggregate level, such as the overall attractiveness, sentiment, or presentation format [8,17,24], often treating short videos as static units. However, what distinguishes short videos from other media formats is their inherent dynamism and sequencing. Adopting a data-driven approach, the frame-by-frame computational analysis enables close examination of how visual compositions evolve over time within each video. On the one hand, this computational approach provides a more accurate and objective way to quantify the presence of specific framing types. Unlike prior approaches that rely on overall ratings or binary labels based on brief cues [10,13], the frame-level method measures how much of a video is actually composed of a given framing style. This leads to a more detailed and representative understanding of framing composition. On the other hand, this approach enables us to assess the frequency of transitions across framing types, a dynamic feature that has been largely overlooked. These transitions capture the complexity and instability of visual presentation and are revealed to influence consumer behavior, and also interact with detailed visual framings. As such, short videos should be understood not only in terms of their content but also in terms of how such content is organized and evolves over time.

Last, this study contributes a replicable and scalable computational approach for analyzing visual framing in short videos. Prior studies have primarily relied on manual coding or survey-based methods to quantify visual features [13,17], approaches that are labor-intensive, subject to human bias, and limited in scale. In contrast, this study integrates deep learning techniques to extract video frames and classify each one according to the defined framing typology. This approach not only allows for fine-grained and temporally sensitive analysis at scale but also offers new possibilities for studying visual framing in other dynamic media contexts. More broadly, it showcases how computational methods can augment theory-driven research, particularly when studying large, multimodal datasets that would be otherwise infeasible to analyze manually.

5.3. Managerial Implications

This study provides practical guidance for improving short-video commerce effectiveness. First, content creators should align visual framing strategies with marketing objectives. Functional and contextual framings are more suitable for supporting product evaluation and sales conversion, whereas social framing should be used cautiously when conversion is the primary goal. The nonlinear findings also suggest that no single framing type should be overused. Second, creators should maintain visual coherence and avoid excessive framing transitions. Although mixed framing types can support storytelling, frequent switching among product-focused, context-focused, and presenter-centered visuals may increase cognitive load and weaken performance. Therefore, similar frames should be grouped where appropriate, and transitions should be used strategically rather than excessively. Finally, marketers should distinguish between engagement-driven and conversion-driven strategies. Presenter-centered visuals may help stimulate interaction and dissemination, while product-focused and context-embedded visuals may be more effective for purchase-related evaluation and sales conversion.

5.4. Limitations and Future Research

There are several limitations that should be addressed in future research. First, this study focuses on book-related short sales videos on Douyin. Although books are suitable for examining how visual framing communicates product-related information, the findings may not fully generalize to other product categories, such as fashion, cosmetics, food, or

electronics. In addition, Douyin has its own platform logic, recommendation mechanism, and user culture. Future studies could examine other product categories and conduct cross-platform comparisons, such as TikTok, YouTube Shorts, or Instagram Reels.

Second, this study relies on observable behavioral outcomes, including shares and sales, rather than directly measuring consumers' cognitive and emotional responses. Future research could use experiments, eye-tracking, physiological measures, or surveys to examine how visual framing affects attention allocation, perceived cognitive load, emotional reactions, and purchase decisions.

Third, our visual framing classification focuses on high-level compositional prominence, namely, whether a frame foregrounds the product itself, the product within a usage context, or the presenter. This operationalization is consistent with our theoretical focus, but it does not capture finer-grained visual features such as facial expressions, camera angles, background complexity, text overlays, product prominence, or aesthetic presentation. Potential classification errors and model bias may also exist, especially when identifying visually diverse product appearances in real-world short-video frames. Future research could further improve the precision and sophistication of visual framing measurement by using more advanced computer vision architectures, emotion AI, aesthetic assessment, and multimodal recognition techniques.

Finally, this study mainly focuses on visual composition and does not fully incorporate other modalities of short-video content. Verbal cues, background music, voice-over style, subtitles, and emotional tone may interact with visual framing to shape consumer responses. Future research could adopt multimodal analysis and cross-cultural validation to examine how visual, textual, auditory, and emotional cues jointly influence engagement and sales in different short-video commerce environments.

Author Contributions: Conceptualization, X.P.; Methodology, X.P., X.X. and L.H.; Validation, L.H.; Formal analysis, X.P.; Resources, X.X.; Data curation, X.X.; Writing—original draft, X.P. and L.H.; Writing—review & editing, L.H.; Visualization, X.X.; Supervision, L.H.; Project administration, X.P.; Funding acquisition, X.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Social Science Fund of China, NSSFC, (22CTQ017). The APC was funded by the same funder.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Restrictions apply to the availability of these data. Publicly accessible video content and engagement metrics were collected from Douyin using web scraping in accordance with the platform's terms of service. Additional sales-related data were obtained from Channama, a third-party social commerce analytics platform. Due to third-party licensing restrictions and platform terms of service, the complete dataset is not publicly available.

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

References

1. Ren, Q. Advertising and content creation on digital content platforms. *Mark. Sci.* **2024**, *43*, 734–750. [[CrossRef](#)]
2. Chen, N.; Yang, Y. The role of influencers in live streaming e-commerce: Influencer trust, attachment, and consumer purchase intention. *J. Theor. Appl. Electron. Commer. Res.* **2023**, *18*, 1601–1618. [[CrossRef](#)]
3. He, D.; Yao, Z.; Teo, T.S.; Ma, Y.; Xu, W. How social learning drives customer engagement in short video commerce: An attitude transfer perspective. *Inf. Manag.* **2024**, *61*, 104018. [[CrossRef](#)]
4. Qin, M.; Qiu, S.; Zhao, Y.; Zhu, W.; Li, S. Graphic or short video? The influence mechanism of UGC types on consumers' purchase intention—Take Xiaohongshu as an example. *Electron. Commer. Res. Appl.* **2024**, *65*, 101402. [[CrossRef](#)]

5. Yang, J.; Zhang, J.; Zhang, Y. Engagement that sells: Influencer video advertising on TikTok. *Mark. Sci.* **2025**, *44*, 247–267. [\[CrossRef\]](#)
6. Zhao, L.; Xu, Y.; Xu, X. The effects of trust and platform innovation characteristics on consumer behaviors in social commerce: A social influence perspective. *Electron. Commer. Res. Appl.* **2023**, *60*, 101284. [\[CrossRef\]](#)
7. Yin, X.; Li, J.; Si, H.; Wu, P. Attention marketing in fragmented entertainment: How advertising embedding influences purchase decision in short-form video apps. *J. Retail. Consum. Serv.* **2024**, *76*, 103572.
8. Xiao, Q.; Huang, W.; Qu, L.; Li, X. The impact of multimodal information features of short sales videos on consumer engagement behavior: A multi-method approach. *J. Retail. Consum. Serv.* **2025**, *82*, 104136.
9. Zhou, B.; Li, J.; Deng, K.; Ling, Z.; Wang, Q. Living an Immersion Experience: Are First-Person Narratives Always Effective for Potential Travelers? *J. Qual. Assur. Hosp. Tour.* **2024**, 1–27. [\[CrossRef\]](#)
10. Tellis, G.J.; MacInnis, D.J.; Tirunillai, S.; Zhang, Y. What drives virality (sharing) of online digital content? The critical role of information, emotion, and brand prominence. *J. Mark.* **2019**, *83*, 1–20. [\[CrossRef\]](#)
11. Araujo, C.J.; Perater, K.A.; Quicho, A.M.; Etrata, A. Influence of tiktok video advertisements on generation z's behavior and purchase intention. *Int. J. Soc. Manag. Stud.* **2022**, *3*, 140–152. [\[CrossRef\]](#)
12. Gu, W.; Chan, K.W.; Kwon, J.; Dhaoui, C.; Septianto, F. Informational vs. emotional B2B firm-generated-content on social media engagement: Computerized visual and textual content analysis. *Ind. Mark. Manag.* **2023**, *112*, 98–112.
13. Wahid, R.; Karjaluoto, H.; Taiminen, K.; Asiati, D.I. Becoming TikTok famous: Strategies for global brands to engage consumers in an emerging market. *J. Int. Mark.* **2023**, *31*, 106–123.
14. Shin, Y.; Noone, B.M.; Robson, S.K.A. An exploration of the effects of photograph content, photograph source, and price on consumers' online travel booking intentions. *J. Travel Res.* **2020**, *59*, 120–139.
15. Gan, J.; Shi, S.; Filieri, R.; Leung, W.K. Short video marketing and travel intentions: The interplay between visual perspective, visual content, and narration appeal. *Tour. Manag.* **2023**, *99*, 104795. [\[CrossRef\]](#)
16. Wang, X.; Zhang, Z.; Jiang, Q. The effectiveness of human vs. AI voice-over in short video advertisements: A cognitive load theory perspective. *J. Retail. Consum. Serv.* **2024**, *81*, 104005. [\[CrossRef\]](#)
17. Meng, L.M.; Kou, S.; Duan, S.; Bie, Y. The impact of content characteristics of Short-Form video ads on consumer purchase Behavior: Evidence from TikTok. *J. Bus. Res.* **2024**, *183*, 114874. [\[CrossRef\]](#)
18. Dong, X.; Liu, H.; Xi, N.; Liao, J.; Yang, Z. Short video marketing: What, when and how short-branded videos facilitate consumer engagement. *Internet Res.* **2024**, *34*, 1104–1128. [\[CrossRef\]](#)
19. Zhao, W.; Hu, F.; Wang, J.; Shu, T.; Xu, Y. A systematic literature review on social commerce: Assessing the past and guiding the future. *Electron. Commer. Res. Appl.* **2023**, *57*, 101219.
20. Lucila, N.B., Jr.; Ching, M.R.D. Understanding social commerce adoption in agriculture: A systematic literature review. *Electron. Commer. Res. Appl.* **2025**, *74*, 101543. [\[CrossRef\]](#)
21. Xiao, L.; Li, X.; Zhang, Y. Exploring the factors influencing consumer engagement behavior regarding short-form video advertising: A big data perspective. *J. Retail. Consum. Serv.* **2023**, *70*, 103170.
22. Peng, Y.; Wang, Y.; Li, J.; Yang, Q. Impact of AI-oriented live-streaming E-commerce service failures on consumer disengagement—Empirical evidence from China. *J. Theor. Appl. Electron. Commer. Res.* **2024**, *19*, 1580–1598.
23. Ge, J.; Sui, Y.; Zhou, X.; Li, G. Effect of short video ads on sales through social media: The role of advertisement content generators. *Int. J. Advert.* **2021**, *40*, 870–896. [\[CrossRef\]](#)
24. Fang, X.; Xie, C.; Yu, J.; Huang, S.; Zhang, J. How do short-form travel videos trigger travel inspiration? Identifying and validating the driving factors. *Tour. Manag. Perspect.* **2023**, *47*, 101128. [\[CrossRef\]](#)
25. Liu, L.; Fang, J.; Yang, L.; Han, L.; Hossin, M.A.; Wen, C. The power of talk: Exploring the effects of streamers' linguistic styles on sales performance in B2B livestreaming commerce. *Inf. Process. Manag.* **2023**, *60*, 103259. [\[CrossRef\]](#)
26. Xiao, L.; Li, X.; Mou, J. Exploring user engagement behavior with short-form video advertising on short-form video platforms: A visual-audio perspective. *Internet Res.* **2026**, *36*, 154–188. [\[CrossRef\]](#)
27. Wang, X.; Wuji, S.; Liu, Y.; Luo, R.; Qiu, C. Study on the impact of recommendation algorithms on user perceived stress and health management behaviour in short video platforms. *Inf. Process. Manag.* **2024**, *61*, 103674. [\[CrossRef\]](#)
28. Entman, R.M. Framing: Toward clarification of a fractured paradigm. *J. Commun.* **1993**, *43*, 51–58. [\[CrossRef\]](#)
29. Meire, M.; Hewett, K.; Ballings, M.; Kumar, V.; Van den Poel, D. The role of marketer-generated content in customer engagement marketing. *J. Mark.* **2019**, *83*, 21–42. [\[CrossRef\]](#)
30. Cheng, M.; Liu, J.; Qi, J.; Wan, F. Differential effects of firm generated content on consumer digital engagement and firm performance: An outside-in perspective. *Ind. Mark. Manag.* **2021**, *98*, 41–58. [\[CrossRef\]](#)
31. Chen, Y.; Sherren, K.; Smit, M.; Lee, K.Y. Using social media images as data in social science research. *New Media Soc.* **2023**, *25*, 849–871.

32. Rahman, M.S.; Sabbir, M.M.; Zhang, J.; Zhang, L. I sense my digital assistants! Assessing the impact of customers' immersive experience and perceived social presence on purchase intention. *Behav. Inf. Technol.* **2025**, 1–19. [\[CrossRef\]](#)
33. Hou, L.; Pan, X. Aesthetics of hotel photos and its impact on consumer engagement: A computer vision approach. *Tour. Manag.* **2023**, *94*, 104653. [\[CrossRef\]](#)
34. Chan, I.C.C.; Chen, Z.; Leung, D. The more the better? Strategizing visual elements in social media marketing. *J. Hosp. Tour. Manag.* **2023**, *54*, 268–289. [\[CrossRef\]](#)
35. Shen, Z.; Pritchard, M.J. Cognitive engagement on social media: A study of the effects of visual cueing in educational videos. *J. Assoc. Inf. Sci. Technol.* **2022**, *73*, 1253–1267. [\[CrossRef\]](#)
36. Kim, K.; Lee, S.; Choi, Y.K. Image proximity in advertising appeals: Spatial distance and product types. *J. Bus. Res.* **2019**, *99*, 490–497. [\[CrossRef\]](#)
37. Li, F.S.; Ma, J. The effect of implied motion in travel photographs on visit intention: The mediating role of mental imagery. *Tour. Manag.* **2024**, *104*, 104919. [\[CrossRef\]](#)
38. Lyu, M.; Huang, Q. Visual elements in advertising enhance odor perception and purchase intention: The role of mental imagery in multi-sensory marketing. *J. Retail. Consum. Serv.* **2024**, *78*, 103752. [\[CrossRef\]](#)
39. Fan, L.; Wang, Y.; Mou, J. Enjoy to read and enjoy to shop: An investigation on the impact of product information presentation on purchase intention in digital content marketing. *J. Retail. Consum. Serv.* **2024**, *76*, 103594.
40. Rietveld, R.; Van Dolen, W.; Mazloom, M.; Worrying, M. What you feel, is what you like influence of message appeals on customer engagement on Instagram. *J. Interact. Mark.* **2020**, *49*, 20–53. [\[CrossRef\]](#)
41. Guitart, I.A.; Stremersch, S. The impact of informational and emotional television ad content on online search and sales. *J. Mark. Res.* **2021**, *58*, 299–320.
42. Chen, S.; Wang, H.; Fang, Y.; Wang, W. Informational and emotional appeals of cover image in crowdfunding platforms and the moderating role of campaign outputs. *Decis. Support Syst.* **2023**, *171*, 113975. [\[CrossRef\]](#)
43. Sweller, J. Cognitive load during problem solving: Effects on learning. *Cogn. Sci.* **1988**, *12*, 257–285. [\[CrossRef\]](#)
44. Van der Land, S.; Schouten, A.P.; Feldberg, F.; van den Hooff, B.; Huysman, M. Lost in space? Cognitive fit and cognitive load in 3D virtual environments. *Comput. Hum. Behav.* **2013**, *29*, 1054–1064. [\[CrossRef\]](#)
45. Li, Z.; Wang, B.; Le Texier, T.; Addai-Dansoh, S. Exploring perception of platform language on consumer decision making: A cognitive load perspective. *J. Electron. Commer. Res.* **2025**, *26*, 50–68.
46. Wang, Y.; Li, K. Not easy to “like”: How does cognitive load influence user engagement in online reviews? *Decis. Support Syst.* **2025**, *192*, 114436. [\[CrossRef\]](#)
47. Wen, T.; Kim, E.; Wu, L.; Dodoo, N.A. Activating persuasion knowledge in native advertising: The influence of cognitive load and disclosure language. *Int. J. Advert.* **2020**, *39*, 74–93.
48. da Silva Cezar, B.G.; Maçada, A.C.G. Cognitive overload, anxiety, cognitive fatigue, avoidance behavior and data literacy in big data environments. *Inf. Process. Manag.* **2023**, *60*, 103482. [\[CrossRef\]](#)
49. Yoo, J.; Kim, J.-H.; Kim, M.; Park, M. Imagery evoking visual and verbal information presentations in mobile commerce: The roles of augmented reality and product review. *J. Res. Interact. Mark.* **2024**, *18*, 182–197.
50. Huang, Z.; Zhu, Y.; Hao, A.; Deng, J. How social presence influences consumer purchase intention in live video commerce: The mediating role of immersive experience and the moderating role of positive emotions. *J. Res. Interact. Mark.* **2023**, *17*, 493–509.
51. Luo, L.; Liu, L.; Zheng, Y.; Chen, J. Visual information and appearance: The impact of visual attributes of user-generated photos on review helpfulness. *Telemat. Inform.* **2024**, *92*, 102164. [\[CrossRef\]](#)
52. Park, J.Y.; Lee, H.E. How Consumer Photo Reviews and Online Platform Types Influence Luxury Hotel Booking Intentions Through Envy. *J. Travel Res.* **2025**, *64*, 1275–1291.
53. Ilicic, J.; Webster, C.M. Eclipsing: When celebrities overshadow the brand. *Psychol. Mark.* **2014**, *31*, 1040–1050. [\[CrossRef\]](#)
54. Brüns, J.D.; Meißner, M. Show me that you are advertising: Visual salience of products attenuates detrimental effects of persuasion knowledge activation in influencer advertising. *Comput. Hum. Behav.* **2023**, *148*, 107891. [\[CrossRef\]](#)
55. Ayres, P. Something old, something new from cognitive load theory. *Comput. Hum. Behav.* **2020**, *113*, 106503. [\[CrossRef\]](#)
56. Sweller, J.; van Merriënboer, J.J.G.; Paas, F.G.W.C. Cognitive architecture and instructional design. *Educ. Psychol. Rev.* **1988**, *10*, 251–296.
57. Wang, Q.; Yang, S.; Liu, M.; Cao, Z.; Ma, Q. An eye-tracking study of website complexity from cognitive load perspective. *Decis. Support Syst.* **2014**, *62*, 1–10. [\[CrossRef\]](#)
58. Fu, S.; Li, H.; Liu, Y.; Pirkkalainen, H.; Salo, M. Social media overload, exhaustion, and use discontinuance: Examining the effects of information overload, system feature overload, and social overload. *Inf. Process. Manag.* **2020**, *57*, 102307. [\[CrossRef\]](#)
59. Lange, C.; Costley, J. The moderating effects of intrinsic load on the relationship between self-regulated effort and germane load. *J. Comput. Assist. Learn.* **2018**, *34*, 652–660. [\[CrossRef\]](#)

60. Willis, J.; Todorov, A. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* **2006**, *17*, 592–598. [\[CrossRef\]](#) [\[PubMed\]](#)
61. Rani, S.; Kumar, M. Social media video summarization using multi-Visual features and Kohnen’s Self Organizing Map. *Inf. Process. Manag.* **2020**, *57*, 102190.

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.