Dec 12th, 12:00 AM

Short-Video Marketing in E-commerce: Analyzing and Predicting Consumer Response

Yutong Guo
National University of Singapore, yutong.guo@u.nus.edu

Chao Ban
Alibaba Group, ban_0330@163.com

Xiao Liu
Alibaba Group, lemon.lx@alibaba-inc.com

Khim Yong Goh
National University of Singapore, gohky@comp.nus.edu.sg

Xixian Peng
Zhejiang University, pengxx@zju.edu.cn

See next page for additional authors

Presenter Information
Yutong Guo, Chao Ban, Xiao Liu, Khim Yong Goh, Xixian Peng, Jiang Yang, and Xiaobo Li

Short-Video Marketing in E-commerce: Analyzing and Predicting Consumer Response

Completed Research Paper

Yutong Guo
National University of Singapore
15 Computing Drive, Singapore
yutong.guo@u.nus.edu

Chao Ban
Alibaba Group
969 West Wenyi Rd, Hangzhou, China
ban_0330@163.com

Xiao Liu
Alibaba Group
969 West Wenyi Rd, Hangzhou, China
lemon.lx@alibaba-inc.com

Khim Yong Goh
National University of Singapore
15 Computing Drive, Singapore
gohky@comp.nus.edu.sg

Xixian Peng
Zhejiang University
866 Yuhangtang Rd, Hangzhou, China
pengxx@zju.edu.cn

Jiang Yang, Xiaobo Li
Alibaba Group
969 West Wenyi Rd, Hangzhou, China
{yangjiang.yj, xiaobo.lixb}@alibaba-inc.com

Abstract

This study analyzes and predicts consumer viewing response to e-commerce short-videos (ESVs). We first construct a large-scale ESV dataset that contains 23,001 ESVs across 40 product categories. The dataset consists of the consumer response label in terms of average viewing durations and human-annotated ESV content attributes. Using the constructed dataset and mixed-effects model, we find that product description, product demonstration, pleasure, and aesthetics are four key determinants of ESV viewing duration. Furthermore, we design a content-based multimodal-multitask framework to predict consumer viewing response to ESVs. We propose the information distillation module to extract the shared, special, and conflicted information from ESV multimodal features. Additionally, we employ a hierarchical multitask classification module to capture feature-level and label-level dependencies. We conduct extensive experiments to evaluate the prediction performance of our proposed framework. Taken together, our paper provides theoretical and methodological contributions to the IS and relevant literature.

Keywords: Short-video marketing, e-commerce, viewing duration, prediction, dataset

Introduction

The recent advances of mobile technologies (e.g., high-quality phone cameras) have drastically reduced barriers to the creation, distribution, and viewing of short-videos. This has boosted an unexpected growth of short-video creators and viewers over the world. By December 2021, the number of short-video users in China has reached 934 million, accounting for 90.5% of the total Internet users (CNNIC 2022). Further, from the U.S., YouTube Shorts, which allows users to create, watch, and share short-form videos, has

achieved 15 billion average daily views over 100 countries by July 2021[1]. Nowadays, given the paucity of consumers' attention across a multitude of media channels, short-videos that are usually 30 seconds or less in duration and contain fast-pace engaging content, outperform other media content types in grabbing audience attention. With these superiorities, short-videos have emerged to be the next frontier of content marketing (Mowat 2018). Particularly, many e-commerce platforms have embraced the short-video-based marketing strategy to promote products on their mobile apps. As of August 2018, the daily-level short-video play count on Taobao, one of the leading e-commerce platforms, is about 1.9 billion[2].

Although prior e-commerce research has demonstrated the positive roles of product videos in influencing consumers' product attitude and purchase behaviors (e.g., Jiang and Benbasat 2007; Kumar and Tan 2015), a limited number of studies documented how content attributes of e-commerce videos influence consumer response. Understanding such a relationship is of great importance for **E**-commerce **S**hort-**V**ideos (ESVs) which we define as *short-videos that are produced by sellers or influencers, published on recommendation pages of e-commerce mobile apps with marketing content for a particular product*. In contrast to videos displayed on products' pages of which a primary goal is to help consumers evaluate product attributes when they have a specific or planned purchase in mind (Kumar and Tan 2015), our focal ESVs are typically intended to inform consumers about new products, stimulate their interests in the advertised products, and enhance their awareness of sellers' or brands' identity when they explore the recommendation pages with less deliberation and unclear purchase plans (Moe 2003). Given these distinct objectives of ESVs, a key concern over ESVs is consumers' avoidance of these informative or persuasive content. When targeted consumers opt to skip over ESVs, the marketers and platforms are likely to lose buyers, traffic, and demand. Hence, producing and recommending ESVs which can evoke positive consumer response are considered as critical goals for e-commerce sellers and platforms. Specifically, we examine ESV consumer response by focusing on viewing behaviors, which is measured as the *average viewing duration or time spent on viewing an ESV*. Our focus is different from recent studies examining ESV effectiveness via consumers' product click and purchase behaviors (Song et al. 2021; Tan et al. 2022). It is suggested that viewing is the very first step in consumers' consumption of marketing content, which is a key indicator of consumers' attention, engagement, and satisfaction (Elpers et al. 2003; Olney et al. 1991; Tong et al. 2020). Thus, ESVs which can receive longer viewing duration are more likely to make consumers interested in the showcased products and ESV creators, thus triggering purchase intent. Further, a better understanding of ESV consumer viewing response can improve ESV recommendation effectiveness and guide ESV content designs. Despite its theoretical and practical importance, consumer viewing response to ESVs has not been fully understood, which thus motivates this study.

Specifically, the first objective of this paper is to draw on the cognitive-affective ad processing framework (MacInnis and Jaworski 1989) to examine how informational and emotional content of ESVs are associated with consumer responses in terms of viewing duration. The cognitive-affective ad processing framework has often been applied to study how informational-emotional marketing content (e.g., ads, marketer-generated posts) affects consumer outcomes, such as brand searching (Chandrasekaran et al. 2018), online purchasing (Guitart and Stremersch 2021), video sharing (Tellis et al. 2019), stop-viewing (Elpers et al. 2003), as well as viewing count (Southgate et al. 2010). Although some prior studies have documented the effects of informational and emotional content of TV commercials or online video ads (Elpers et al. 2003; Guitart and Stremersch 2021; Tellis et al. 2019) which are similar with ESVs, they have neither distinguished specific attributes of informational and emotional content nor related those attributes to the outcome of consumers' viewing duration in the context of e-commerce. Moreover, when using e-commerce mobile apps for shopping, consumers often have specific purchasing goals (Lerche et al. 2016) and they will intrinsically want to seek more product information to reduce purchase uncertainty (Pavlou et al. 2007). Thus, consumers' ad processing of ESVs in e-commerce may inherently differ from that of other video ad content in traditional media. Given the distinctiveness of our research context, prior findings may not be directly generalized to predict consumers' viewing behaviors of ESVs. Hence, our study aims to fill in this gap by relating ESV consumers' viewing response with ESV informational-emotional content attributes.

Our second research objective, in addition to investigating the association between ESV content attributes and consumer response, is to predict consumer response to ESVs. This is a critical challenge faced by e-commerce platforms which rely on consumer response predictions to enhance ESV recommendation

[1] https://earthweb.com/youtube-shorts-statistics/
[2] https://www.alibabagroup.com/en/ir/presentations/Investor_Day_2018_Taobao.pdf

efficiency, given that ESV content quality is highly heterogenous due to discrepant production resources, recording skills, and product categories. Recent advances in machine learning (ML) have provided the methodological foundation for such consumer response predictions. However, there still exist at least two challenges for this task. First, there are no existing video datasets specific to ESVs. The available datasets for relevant predictions (e.g., viewing duration, video popularity) are mainly constituted by user-generated videos (UGVs) on video sharing platforms (Demarty et al. 2017; Park et al. 2016; Wu et al. 2018). ESVs significantly differ from UGVs in terms of content creators (i.e., brands and sellers vs. online users), topics (i.e., product-centric vs. entertainment-oriented), durations (i.e., short vs. long-format) etc. To some degree, the datasets that were built to understand video ads effectiveness are related to our work (Hussain et al. 2017). However, these datasets generally have a small sample size and do not contain labels for viewing duration predictions. Second, in the face of the cold-start setup in e-commerce content marketing such that high volumes of short-videos are created but few historical user-level interaction data is available, predicting consumer response to ESVs becomes more challenging. One potential solution is to leverage on video content, but the prevailing characteristics of ESVs can lead to more complex multimodal information representations. For instance, ESV creators often resort to clickbaits – a kind of eye-catching but irrelevant information (i.e., titles) to get more eyeballs (see Figure 1a). Hence, how to extract multimodal features of ESVs to simulate information processing in the human brain and obtain more robust representations is an important issue for ESV consumer response predictions.

To achieve the aforementioned research objectives, this paper draws on insights from multiple disciplines including information systems, marketing, and computer science. Based on prior research on e-commerce videos (Flavián et al. 2017) and ad processing framework (MacKenzie et al. 1986; Maclnnis and Jaworski 1989), we first propose four informational (i.e., product description and demonstration) and emotional (i.e., pleasure and aesthetics) ESV content attributes and theorize how they affect consumer viewing response toward ESVs. Following the best practices in the computer vision literature, we then construct a large-scale e-commerce short-video dataset (ESVD) from Taobao, which contains in total 23,001 ESVs across 40 product categories, including clothing, home appliances, snacks, etc. To construct the consumer response label, we collect the total playtime and view count over 7 days after an ESV is published on Taobao. Additionally, we recruit a group of annotators to evaluate informational and emotional content attributes. The dataset will be publicly available upon this paper's publication. We hope the availability of this ESV dataset can provide the impetus for future research.

Using the constructed dataset, we adopt a mixed-effects linear regression model to quantify the relationship between ESV content attributes and consumer response. We find heterogeneous effects of particular ESV content attributes on increasing consumers' viewing duration of ESVs across different ESV types by purchase behavior categories. Additional results also suggest that the proposed content attributes are critical factors of ESV effectiveness in terms of consumers' interactions and purchases. Further, to predict consumer response to ESVs, we propose a content-based multimodal multitask framework termed IDiGM^2. Specifically, we first propose the information distillation module to distinguish between and extract the shared, special, and conflicted information from different modality features (illustrated in Figure 1a), in order to obtain a comprehensive representation of ESVs. Second, based on the relations between ESV content attributes and consumer response, we employ a hierarchical scheme to jointly learn two sub-tasks – predictions of ESV content attributes and consumer viewing duration. We utilize the sub-task dependencies in terms of both feature representations and labels in the multitask learning. As we show, our proposed framework results in more comparable and consistent ESV consumer response prediction results.

The ability to understand and predict ESV consumer response has many practical implications. Our results on the contributing sources of ESV consumer response can guide creators to better design and produce informational and emotional ESV content across product categories. For instance, to create ESV promoting apparel products, creators should attach more importance to informational content but put less focus to aesthetics-related content. Moreover, our work can help e-commerce platforms to efficiently manage and distribute a large number of short-videos. E-commerce platforms can rely on our proposed prediction framework of ESV viewing duration to pick out attractive ESVs which describe product attributes information, demonstrate how to use the showcased products, arouse a delighted feeling, or are visually pleasing to increase platforms' content marketing and recommendation effectiveness.

The main contributions of this paper are summarized as follows. First, our study is among the pioneers to examine the effectiveness of short-video marketing in the e-commerce context based on the ad cognitive-

affective processing framework. By quantifying the roles of ESV content attributes on consumer response to ESVs, our paper extends the IS and marketing literatures on the effect of informational and emotional ad content (e.g., Guitart and Stremersch 2021; Tellis et al. 2019). Second, we propose a novel multimodal multitask framework to predict ESV consumer response with content attributes by efficiently extracting information from ESVs' visual frames, audio information, and associated titles. Our proposed framework can enable scalable video content analytics and viewership response predictions. Third, this study provides both dataset and methodological contributions to the IS and marketing literatures on visual data analytics (Liu et al. 2020; Zhou et al. 2021). Specifically, we construct a pioneering, to-be publicly available large-scale, diverse-products ESV dataset containing a rich set of annotated ESV content attributes and consumer response labels in terms of average viewing durations.

Notes: The red boxes show the modality-shared information, the blue boxes show the modality-special information, and the green boxes show the modality-conflicted one.

Figure 1a. Illustration of Extracted Information from the Visual and Textual Modality of an ESV

Figure 1b. A Screenshot of an ESV Display Page

Related Work

Consumer Response to Online Advertising Videos

Considering the advertising nature of ESVs, our paper is related to various prior research below examining how video attributes affect online ad effectiveness in terms of consumer response. First, in the context of e-commerce, although most prior IS and marketing research has uncovered the overall positive role of videos on product webpages in driving consumer purchases (Jiang and Benbasat 2007; Kumar and Tan 2015), there exist a limited number of studies documenting that such a positive effect is due to specific video attributes. For example, Flavián et al. (2017) has suggested that the content type of product videos in e-commerce environments determines consumer response such that demonstration-oriented product videos are effective for influencing consumer attitude and purchase intentions but promotional-centric ones are not. Moreover, a recent work by Song (2021) has revealed that static and dynamic visual features of e-commerce advertising videos affect consumers' click behaviors for the advertised products. Second, beyond the context of e-commerce, prior literature has been much devoted to understanding the factors of advertising video performance on social media platforms, such as YouTube, Facebook, and TikTok. Generally, they studied how consumers' viewing, sharing, and buying behaviors are affected by content and format features of advertising videos (e.g., Southgate et al. 2010; Tellis et al. 2019; Yang et al. 2021). For instance, Teixeira et al. (2012) concluded that videos which can evoke joy and surprise engage viewers longer and more attentively. In addition, prior work also found that vertical mobile videos compared with horizontal ones can reduce viewing efforts and increase processing fluency as well as viewing duration (Mulier 2021). Despite these significant efforts, no previous studies have examined how informational and emotional content attributes are related to consumers' viewing responses to ESVs.

The Roles of Informational and Emotional Content

In this section, we draw on the insights from the advertising literature to propose potentially important content attributes that influence viewers' viewing behaviors. Prior literature has proposed a two-pathway framework that advertisements influence consumer attitude toward the ad through the cognitive and affective processing routes (MacKenzie et al. 1986; Maclnnis and Jaworski 1989). Based on this ad processing framework, a number of empirical research has explored how informational and emotional ad content drive consumer response (Elpers et al. 2003; Tellis et al. 2019). Informational ad content refers to the content conveying the factual information about the features and/or benefits of the advertised products while emotional ad content highlights the non-factual content that evokes favorable feelings and emotions (MacInnis et al. 2002). Prior research has shown the differential roles of informational and emotional ad content on ad effectiveness in terms of different aspects of consumer behaviors. Specifically, it has been found that informational ad content can induce consumers to stop viewing TV ads (Elpers et al. 2003) and reduce online sharing (Tellis et al. 2019) but increase brand searching (Chandrasekaran et al. 2018) and product purchasing (Guitart and Stremersch 2021). Further, most prior studies have found a positive effect of emotional ad content on consumer responses (e.g., viewing (Elpers et al. 2003), sharing (Tellis et al. 2019), purchasing (Guitart and Stremersch 2021)), whereas insignificant effects were also reported (Chandrasekaran et al. 2018). Our paper differs from these prior studies in two important aspects. First, in contrast to prior work on TV or social media ad content, we focus on the marketing content (i.e., ESVs) in e-commerce where consumers tend to have higher intentions of acquiring product-related information. Second, instead of differentiating only informational-emotional dimensions, we further distinguish four specific content attributes so that we can compare the relative importance of these attributes and provide more concrete insights for researchers and marketers.

Based on the marketing literature, we propose two informational content attributes which are product description and product demonstration. Like many product ads, ESVs are designed to describe and explain a product's detailed features and benefits or functional advantages. Previous research has consistently suggested that factual descriptions of product attributes in advertising videos are important drivers of consumers' product interests and purchase intentions (Guitart and Stremersch 2021; Kumar and Tan 2015). This is because consumers are tempted to seek more product information to reduce product uncertainty in purchase decision processes. In addition, to promote new or niche products, sellers often leverage ESVs to demonstrate how products work in specific scenes to reduce consumer uncertainty. Previous studies have shown that demonstrating products in concrete consumption scenarios is also an effective way to enhance consumers' perceived product tangibility and direct experience (Flavián et al. 2017). Thus, product demonstration in ESVs that can allow consumers to have a more realistic impression of products (Flavián et al. 2017) may sustain consumers' viewing. In sum, both product description and product demonstration in ESVs may exert critical influence on consumer viewing response to ESVs.

In addition to informational content attributes, we put forward two emotional content attributes which are pleasure and aesthetics. Pleasant content in video ads or movie trailers is suggested to play an essential role in boosting viewing intentions (X. Liu et al. 2018) and conversions in terms of shares and sales (Guitart and Stremersch 2021; Tellis et al. 2019). Similar to conventional marketing videos, we conjecture that ESVs could also be strategically produced to influence viewers' pleasant feeling and emotions that could have beneficial effects on viewing behaviors. As such, pleasure content attribute should be considered as an important aspect of emotional content attributes. In addition, research on consumer psychology has consistently highlighted the value of aesthetics on consumers' feelings in various marketing contexts, such as retailing environment (Vieira 2010), website design (Lavie and Tractinsky 2004), and product evaluation (Hagtvedt and Patrick 2008). Generally, consumers tend to prefer beautiful or visually-pleasing objects, people, or consumption environments (Hoyer and Stokburger-Sauer 2012). As such, ESVs, which involve products and people in specific scenes, can also rely on aesthetic cues to stimulate viewers' positive feelings in order to achieve favorable evaluations of the ESVs. As such, aesthetics can serve as another key ESV emotional content attribute to influence consumers' viewing response.

Overall, the four proposed ESV content attributes are not mutually exclusive; one ESV can leverage on multiple attributes, though resource constraints may discourage extensive capitalization of all attributes. Thus, it is necessary to evaluate the relative importance of the effects of ESV content attributes on consumer response. Furthermore, ESVs are created for marketing an extensive variety of products. When making purchase decisions on different types of products, consumers usually show different levels of information

processing and involvement (Munthiu 2009), which may moderate the response effect of specific content attributes. Thus, we also study how the effects are contingent on different purchase behavior categories.

Video Datasets

To the best of our knowledge, no public datasets are available to predict the average viewing duration of ESVs. However, there are several well-established datasets which are related to our work. For instance, Xie et al. (2020) collected view counts for 11,219 micro-videos to predict video popularity. Other two datasets which have labels of viewing duration were developed for predicting viewer engagements on YouTube (Park et al. 2016; Wu et al. 2018). Other annotated video datasets, with labels of creativity (Redi et al. 2014), interestingness (Demarty et al. 2017), and memorability (Cohendet et al. 2018), describe the subjective perceptions of video quality. Although consumer response is conceptually related to the labels provided by prior studies, consumers' preferences may vary when viewing ESVs on e-commerce platforms versus viewing entertainment-oriented videos (e.g., movies) collected by the extant datasets. Other relevant datasets were developed to understand video ads, providing labels of interestingness, topics, sentiments, effectiveness, and climax (Hussain et al. 2017; Jiang et al. 2013; Vedula et al. 2017). Video ads and ESVs have similar purposes in terms of communicating essential product information with consumers. However, unlike video ads, ESVs are typically created mostly by non-professionals. Moreover, given the low cost and high ease of producing and publishing ESVs, many sellers or influencers on e-commerce platforms can easily promote their products through ESVs. Hence, the types of marketed products in ESVs are much more diverse than those of traditional video ads. In addition, the existing video ads dataset has a rather small sample size, which is limited for training robust prediction algorithms.

Overall, we conclude that it is necessary and critical at this juncture to develop a new benchmarking dataset for analyzing consumers' viewing response to ESVs. This new dataset aims to have a relatively large-scale, reliable set of ground-truth labels and to consist of a diverse set of marketed products in the ESVs.

Predicting Consumers' Viewing Response

Our work is related to studies predicting video viewing durations. Park et al. (2016) found that popularity metrics such as likes and comments have a significant predictive power for viewing duration of YouTube videos. However, such predictions may not be valid for cold-start videos which have few user interactions. Following Park et al. (2016), Wu et al. (2018) used a set of video meta-data, such as duration, language, topics, and channel reputation features to predict the proportion of YouTube video watched with linear regressions. Neither work has extracted more sophisticated features of multimodal video content, which is the focus of our proposed prediction framework. Since multimodal features could mimic human understanding of multimedia, multimodal learning methods have been proposed for predicting relevant video properties. For example, Jiang et al. (2013) linearly fused kernels computed from visual, audio and high-level attribute features to predict video interestingness. Chen et al. (2018) developed multimodal fusion strategies at different levels for predicting the total number of views for TV serials. To assess video aesthetic quality, Kuang et al. (2020) proposed a multi-stream learning framework that consists of spatial, motion, and structural streams to capture the multimodal features. As such, predicting consumer response to ESVs with a multimodal learning framework could be a potential solution for our objective.

Since content attributes used to predict consumer response are highly correlated, some studies on video content understanding have employed a multitask learning approach, which gives a better prediction performance by enabling each task to help others. Ye et al. (2018) designed a multitask learning framework to predict topics and emotions in video ads. Liu et al. (2018) developed a multitask perceptual embedding method for video emotion and interestingness predictions. However, they did not consider the hierarchical relationships between the specified multiple tasks. In this paper, we utilize a new hierarchical multitask classification method which considers feature-level and label-level dependencies.

ESV Dataset Construction

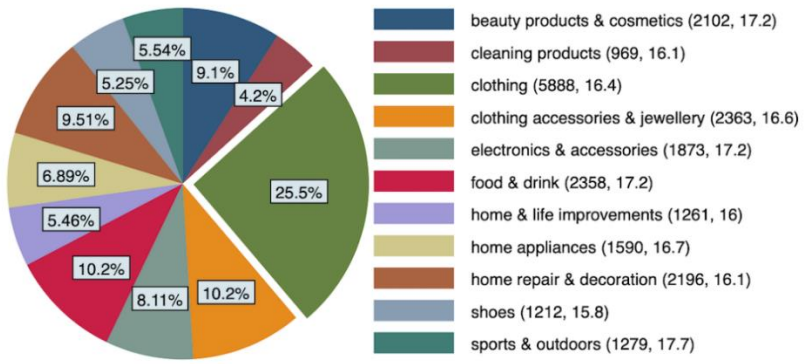
To address the challenges of the analysis and prediction of ESV consumer response, this study first aims to construct a large-scale ESV dataset. In this section, we detail how we construct the dataset, including the video collection strategy, consumer response measures, and ESV content attributes annotations.

Collecting ESVs

We collaborated with and gathered ESVs from a leading global e-commerce platform based in China – Taobao. Taobao users are exposed to the ESVs when they browse its mobile app's homepage, product main pages, and sellers' or influencers' (called "Daren" on Taobao) main pages. The ESVs are presented with titles, creator information, and the showcased product's link in full screen mode (see Figure 1b).

We retrieved data from Taobao's internal database and obtained the properties for ESVs that were published on Taobao homepage recommendation pages between 14 and 21 November in 2020. To control for the potential confounding effects of video duration, exposure count, and showcased product category, we removed some ESVs. Specifically, to control the effect of ESV duration on consumers' viewing duration, we retained only the ESVs which were between 10 to 30 seconds in duration. To analyze only ESVs with some minimal viewing responses, we kept ESVs that had at least 50 exposures one day after their publication. To focus primarily on the dominant types of ESVs, we included ESVs with showcased products that were from the top-40 product categories in terms of the number of ESVs, covering about 85% of ESVs on Taobao. Our sample ESVs feature a number of products that were seldom shown in video ads, such as makeup remover pads, phone cases, and wallpapers. Overall, this ESV data sampling approach garnered 156,097 ESVs of 96,130 unique products in the initial dataset.

In order to manage the content attributes annotation task more efficiently later (see subsection "Annotating ESV Content Attributes"), we further used a stratified sampling strategy to select a subset of ESVs from the initial one described above. We first manually aggregated 40 product categories of ESVs into 11 ESV categories (shown in Figure 2). Each ESV category is then considered as a stratum. According to the proportion of each stratum in terms of the number of ESVs, simple random sampling was applied within each stratum. We also conducted a de-duplication process to remove similar ESVs based on the similarity of visual frames. Our final ESVD then contains 23,001 ESVs from 10,083 unique creators and spans 16.6 seconds on average across all ESVs. Detailed information of the ESVs by category is shown in Figure 2.



Note: For each ESV category, the number of ESVs and average duration are given in the brackets.

Figure 2. Breakdown of ESVs by Category

Measuring Consumer Response to ESVs

Since consumers can easily swipe-up to the next ESV if the current one is not appealing, the average viewing duration of consumers can be an important indicator of viewers' responses to ESVs. We collected the two related daily-level measures for each ESV: number of views and total viewing duration across 7 days after its release on Taobao's mobile app. We aggregate the daily-level measures into the weekly average viewing duration to measure the consumer response to ESVs because some ESVs on Taobao may only capture a significant number of views after a longer period of time. Wu et al. (2018) proposed a relative viewing duration measure (i.e., weighted by video duration) in the context of YouTube offering either long-form or short-form videos. However, for our ESV data, as the duration is constrained within a short and fairly uniform time-interval (i.e., 10-30 seconds, mean = 16.6 and standard deviation = 5.0), we believe that using the absolute viewing duration to measure ESV's consumer response is reasonable.

Annotating ESV Content Attributes

To construct the labels for our proposed four ESV content attributes, we first adapted the annotation questions from prior published marketing studies (Bruce et al. 2020; Tellis et al. 2019). Annotators coded the questions with a binary scale (*Yes, I agree/ No, I do not agree*). Before the formal annotations, we conducted two rounds of pilot test through a survey platform (https://www.wjx.cn) to examine whether the annotation questions were clear and understandable. Volunteers were asked to watch five ESVs and evaluate the content attributes. Moreover, they were instructed to offer at least a suggestion on improving the annotation questions we provided. We revised the question phrases after each round of pilot test accordingly. Our final annotation questions are shown in Table 1.

Table 1. Annotated ESV Content Attributes and Questions

ESV Content Dimension	*ESV Content Attribute*	*Annotation Question*	*Number of ESVs Coded as 1*
Informational	Product description	The ESV describes the specific features, functions, or benefits of the showcased product.	12,175 (52.93%)
	Product demonstration	The ESV demonstrates how to use the showcased product in a specific consumption scenario.	8,037 (34.94%)
Emotional	Pleasure	The ESV gives you a pleasurable and delighted feeling.	5,327 (23.16%)
	Aesthetics	The ESV is visually pleasing and comfortable to view during its playback.	12,251 (53.26%)

ESV content attributes were formally labelled on the Alibaba internal Data Service platform (ADS)[3]. To ensure high-quality annotations, the ADS platform enables an "automatic" check where requestors can provide a subset of pre-labelled tasks, allowing automatic comparisons between the annotated labels and the true labels. With this mechanism, we are able to select and recruit only the experienced, competent annotators (with overall task acceptance rate of higher than 98%). Moreover, the ADS allows the annotators to receive feedback and re-annotate the incorrect annotations immediately during the annotation process.

Each ESV was labelled by 5 independent annotators. In total, 25 paid annotators (12 male and 13 female) were recruited, carefully trained and completed the annotation tasks from 26 November to 10 December in 2020 on ADS[4]. Each individual annotator assessed between 4,500 and 4,700 ESVs. During the annotation process, the annotators were advised to listen to the audio and watch an ESV completely. No time constraints were set. We relied on the "majority wins" strategy to determine the final labels. Specifically, we considered each ESV content attribute that is labelled as positive by a majority of annotators (i.e., at least three) to be truly positive as such. We note that 81% of ESVs annotated show full agreement by the annotators on the pleasure attribute, which has the highest 100% agreement among the four attributes. The proportions of ESVs that show full agreement on the other three attributes are between 72% and 78%, indicating that the reliability of the annotated labels is generally high. The frequency statistics of the four annotated attributes are shown in Table 1's last column. We find that the product description and aesthetics are the most common two content attributes while pleasure is less observed in our sampled ESVs.

In addition to the four content-related attributes, annotators were asked to answer three additional quality-control questions: (1) "*Does this ESV have background music? (Yes/ No)*", (2) "*Does this ESV have human voices/ narrations? (Yes/ No)*", (3) "*Considering the showcased products, which product category do you think this ESV should be?*". For the first two quality-control questions that are expected to have consistent answers from different annotators, we calculated the inter-annotator agreement and ask the "minority" group of annotators to re-annotate after an ESV received 5 labels. Such quality control intervention was conducted every day during the annotation process. For the third question, we provided annotators with the 11 aggregated ESV categories from which they could select only one or the alternative "*not clear*". After comparing the true labels and the annotated labels of ESV categories, we obtained an overall accuracy rate of 93% for ESV category judgement, indicating that our annotated dataset has a

[3] The ADS offers high-quality outsourced data annotation services which can be customized task by task. The task requestors can communicate with the annotator teams directly for quality control.

[4] We provided detailed annotation instructions with specific examples to all the annotators. For each annotation question, we provided specific examples.

high quality. About 8% of ESVs received the label "not clear", suggesting that the majority of ESVs have a clear content topic spanning distinct product categories.

Analysis of ESV Content Attributes

To answer how ESV content attributes drive consumer response in terms of average viewing durations, we specify and estimate a two-level mixed-effects linear model to account for both ESV-level and creator-level unobserved characteristics. Our focal dependent variable, *ViewDur,* is measured by an ESV's average viewing duration over 7 days after its public release on Taobao[5]. Our four focal independent variables of ESV content attributes (i.e., *ProdDesc*, *ProdDemo*, *Pleasure*, *Aesthetics*) are measured according to human annotations. Besides, we include several control variables to capture both ESVs' and ESV creators' key properties. To account for audio effects, we include two dummy variables (i.e., *Music* and *Narration*) according to the annotation results. We also accounted for the format of ESVs in terms of height and width. Specifically, *Horizontal* is used as the baseline and the other two formats (i.e., *Squared* and *Vertical*) are included as dummies in the regression model. To account for the effect of human presence, we introduce a control variable (i.e., *Faces*) that indicates whether human faces are shown in ESVs. The duration of an ESV (i.e., *Duration*), the total number of views (i.e., *ViewCount*) and interactions (i.e., likes, comments, shares) (i.e., *Interaction*) an ESV obtains over the 7-day post-release period, ESV publication date dummies, and 11 ESV category dummies are also included as controls. Moreover, we control for creator effects by introducing ESV creator type dummies (i.e., *Seller* and *Daren*[6]) and creators' historical performance in terms of published content (i.e., ESVs and images) view count in the past week (i.e., *PastWeekViews*). To assess multicollinearity, we run ordinary least squares regression models and obtain variance inflation factors (VIF) for each variable. The VIF values are all below 4, indicating that multicollinearity is not a concern in our analysis[7].

The main model estimation results are reported in Model 1 in Table 2. Our proposed ESV content attributes have positive effects on the average viewing duration of an ESV. Among the four attributes, product demonstration (0.565, p<0.001) exerts the greatest effect on consumer viewing duration, followed by product description (0.380, p<0.001). The results imply that an ESV that demonstrates how to use a product or delivers more factual descriptions about the product can attract significant viewers' attention in terms of the viewing duration. For the emotional content attributes, the associated coefficients for *Pleasure* and *Aesthetics* are 0.170 (p<0.001) and 0.196 (p<0.001) respectively, suggesting that an ESV that evokes viewers' pleasant feeling and is aesthetically pleasing is associated with longer average viewing durations.

To further evaluate the business performance of the proposed ESV content attributes, we examine how they are associated with two additional dependent variables which are *Interaction* and the number of orders for the showcased product as placed by an ESV's viewers (i.e., *Order*). The former indicates customer engagement with ESVs and the latter measures the economic values of ESVs. The model estimation results are shown in Model 2 and 3 in Table 2. On the engagement effect, we find significant coefficients for *ProdDesc* and *Pleasure* but with opposite signs, which is consistent with prior findings on the negative role of informational content and positive role of emotional content on YouTube video ad sharing (Tellis et al. 2019). ESVs delivering factual descriptions about products may be more argumentative and less interesting, and thus lower consumers' interaction intensity. Additionally, on the purchase outcome, we find that product demonstration and aesthetics of ESVs are both driving forces of product orders. However, pleasurable content is found to diminish the demand of showcased products. One plausible explanation is that consumers are likely to exert less mental effort to process peripheral cues such as positive emotions in advertising content (Chandrasekaran et al. 2018), which may inhibit consumers' motivation and ability to process and evaluate the showcased product itself, and thus discouraging purchase behaviors.

In addition to the full-sample analysis, we conduct a set of subsample analyses in order to understand whether the relationships we observe vary across different types of ESVs. Given that consumers' purchase behaviors may vary across different types of products, we aggregate ESVs into four main types based on

[5] Using the relative viewing duration measure, we obtain consistent results in terms of the coefficient sign and significance. These results are available upon request.

[6] The baseline type is normal Taobao users who also generate ESVs. The proportion of user-generated ESVs in our dataset is about 10%.

[7] Control variables of count measures are specified in the logarithmic form for skewness control.

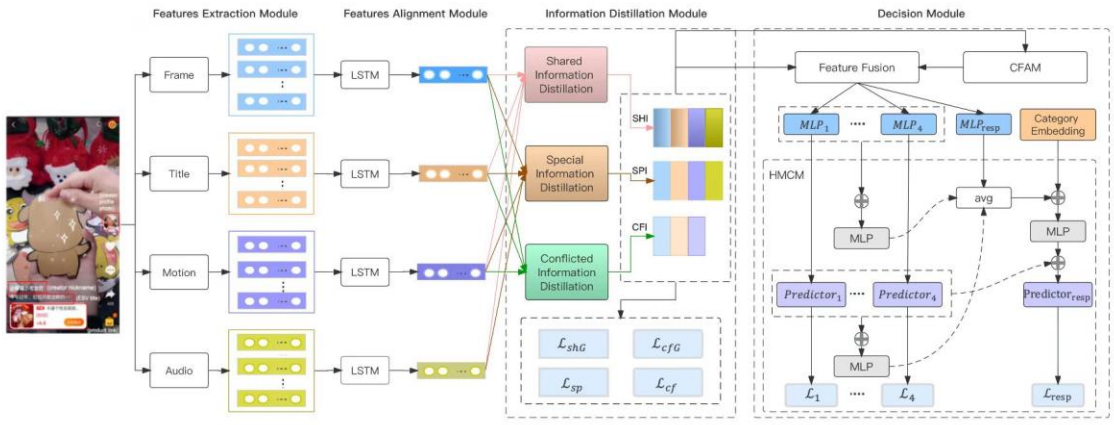
how the showcased products in ESVs would elicit different purchase behaviors: complex (e.g., major household appliances), dissonance-reducing (e.g., jewelry), habitual (e.g., fresh food), and variety-seeking (e.g., clothing) purchases (Kotler and Armstrong 2014). We estimate the models for each type of ESVs by this classification. As shown in Table 2, Model 4-7, the associated effects of the four content attributes are heterogeneous across different ESV types by purchase behavior categories. For instance, the informational content of ESVs promoting products which elicit complex purchases can drive longer viewing duration while emotional content has no effect on viewing response. This could be because products associated with complex purchases are usually expensive, hardly understood, and induce higher purchase risks, which could lead consumers to be highly involved in seeking product attribute information from various channels, such as ESVs, to make purchase decisions. These content heterogenous effects suggest that ESV creators should consider product characteristics and consumer purchase behaviors when producing ESV content.

Table 2. Mixed-Effect Regression Model Estimation Results

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7
	ViewDur	ln*Inter-action*	ln*Orders*	*ViewDur*			
				Complex	*DissonanceR*	*Habitual*	*Variety*
ProdDesc	0.380***	-0.039**	0.017	0.296**	-0.084	-0.133	0.773***
ProdDemo	0.565***	0.017	0.025**	0.446***	1.074***	0.779***	0.316***
Pleasure	0.170***	0.077***	-0.027**	-0.048	-0.099	0.305***	0.149**
Aesthetics	0.196***	0.001	0.025**	0.075	0.398***	0.348***	0.093
Music	-0.289***	0.005	-0.128***	-0.171	-0.086	-0.359***	-0.324***
Narration	0.114**	0.121***	0.014	0.379**	0.110	-0.033	0.176**
Squared	0.598***	0.095**	0.047	0.232	1.545***	1.184***	0.190
Vertical	0.241**	0.087**	0.053**	-0.001	0.709***	0.113	0.163
Faces	-0.037	0.103***	-0.016	0.129	-0.181*	-0.099	0.040
Duration	0.095***	-0.017***	-0.008***	0.086***	0.079***	0.117***	0.093***
ESVs	23,001	23,001	23,001	2,081	3,084	5,435	12,401
Creators	10,083	10,083	10,083	892	1,920	2,848	6,003
R^2 (ESV)	0.175	0.196	0.178	0.170	0.185	0.152	0.129
R^2 (Creator)	0.201	0.223	0.193	0.205	0.202	0.168	0.142

Notes: *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Standard errors are omitted for the sake of brevity. All control variables, ESV publishing date dummies and ESV category dummies are included in the models.

Predicting Consumer Response to ESVs



Notes: Multimodal features are firstly extracted and aligned to the same dimension space via the Feature Extraction and Feature Alignment Modules in sequence. The feature vectors are then fed into the Information Distillation Module to extract shared, special, and conflicted information. Such high-level information is sent to the Decision Module to hierarchically predict ESV content attributes and consumer responses. $\mathcal{L}_1$ – $\mathcal{L}_4$ are the loss for the ESV content attributes. $\mathcal{L}_{\text{resp}}$ is the loss for ESV consumer response.

Figure 3. IDiGM^2 Framework Architecture

The Information-DIstillation Guided Multimodal Multitask Framework (IDiGM²)

In this section, we present the proposed Information-DIstillation Guided Multimodal Multitask framework (*IDiGM²*) for predicting ESV consumer response by utilizing ESVs' multimodal content features (see Figure 3). Next, we evaluate the performance of our proposed framework with a set of experiments on ESVD.

The overall architecture of our proposed multimodal multitask framework is shown in Figure 3. Following Ma et al. (2019), we divide the range of ESV average viewing duration into three intervals (i.e., short, medium and long viewing duration) and treat the ESV consumer response prediction as a three-class classification problem. In sum, the proposed framework contains four modules: Feature Extraction Module, Feature Alignment Module, Information Distillation Module, and Decision Module, as described below.

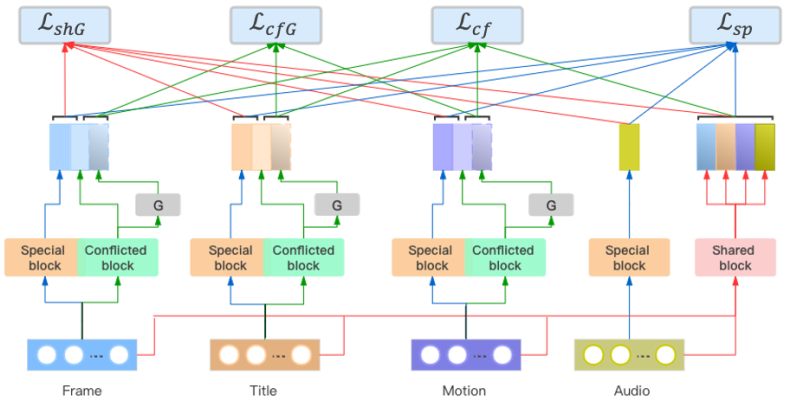
Feature Extraction Module (FEM)

This module is to extract ESV multimodal features with several pre-trained models of which the accuracy can reach over 90%. Four multimodal features are extracted from three ESV modalities, namely images, audios, and titles. Specifically, for images, we rely on the Inception-Resnet V2 (Szegedy et al. 2016) and the S3D network (Miech et al. 2020) to extract frame and motion features, respectively. Moreover, we extract audio features through the VGGish network[8]. Last, we employ the BERT-style model (Devlin et al. 2018) to obtain ESV title embeddings. The four feature matrices are sent into the second module for alignment.

Feature Alignment Module (FAM)

The purpose of this module is to fuse the sequential information and align the feature dimensions. Each modal feature vector is fed into its own LSTM network to fuse the sequential information as well as align the feature vector dimension so that the inputs of the next module is in the same dimension space.

Information Distillation Module (IDM)



Notes: *G* is the generative network. We use different colours to distinguish three types of information streams, namely, red: *SHI*, blue: *SPI*, and green: *CFI*.

Figure 4. IDM Framework Architecture

Inspired by the multimodal information processing style in the human brain, we design the IDM (illustrated in Figure 4). Due to either intended or inexperienced ESV content designs, the cross-modality information uniqueness and/or discrepancy are common in ESVs (see Figure 1a). When humans encounter modality-shared (*SHI*), modality-special (*SPI*), and modality-conflicted (*CFI*) information, they tend to have different information processing modes. Specifically, humans are more likely to trust the shared information but can spend more time on evaluating the special information that is unique in a single modality. Moreover, different modalities may deliver contradictory information for the same characteristic of an object, people,

[8] https://github.com/tensorflow/models/tree/master/research/audioset/vggish

or scene. For this kind of conflicted information, humans may select to trust one of the sources or directly abandon them. In order to capture the way that humans process the ESV content, the IDM is designed to distinguish and extract the aforementioned three types of information from low-level modality features. Correspondingly, the module is composed of three sub-modules which are detailed below.

SHared Information Distillation Module (SHID). This sub-module is used to extract shared information from the four modality feature vectors of an ESV's image frame, motion, audio and title. To achieve this goal, we use the generative adversarial nets (GANs)-style method (Goodfellow et al. 2014). The generator targets to recover the original information for a single modality based on its special and conflicted information together with the SHI from other modalities. The discriminator aims to distinguish between the generated modality features and the original ones. Such a min-max game will enable the SHID to obtain the *SHI* among the four modal features. Guided by GANs, the loss function of the SHID is specified as below:

$$\begin{aligned} X_G &= G(X_{SPI}, X_{CFI}, Y_{SHI}) \\ \mathcal{L}_{shG} &= -\mathbb{E}(D(X_G)) \\ \mathcal{L}_{shD} &= \mathbb{E}(D(X_G)) - \mathbb{E}(D(X_R)) \end{aligned} \tag{1}$$

where G is the generative network; D is the discriminative network; X_{SPI} and X_{CFI} are the *SPI* and *CFI* of modality X; Y_{SHI} is the *SHI* from any other modality Y; X_G is the generated vector of modality X; X_R is the original vector of modality X; $\mathcal{L}_{shG}$ and $\mathcal{L}_{shD}$ are the generative and discriminative network loss.

SPecial Information Distillation Module (SPID). Since each modality may contain its unique information, we design the SPID to extract such modality-special information. Such information is specific to one of modalities and is not related to information from other modalities. Thus, the special information representation of each modality should be irrelevant to its own shared and conflicted information and all kinds of information of other modalities. In other words, the cosine distance between *SPI* and all other information vectors should be as large as possible. Hence, the loss function of the SPID is constructed as:

$$\mathcal{L}_{sp} = abs(\cos(X_{SPI}, Y)) \tag{2}$$

where X_{SPI} is the *SPI* of modality X; Y is the set of all information vectors except X_{SPI}.

ConFlicted Information Distillation Module (CFID). Since the CFID is designed to extract the conflicted information between two modality pairs (i.e., title vs. frame and title vs. motion), the output vectors of the CFID for each pair of modalities should be contradictory. As such, the conflicted information of one modality (e.g., title) could be projected to a vector space which should be close to the conflicted information of the paired modality (e.g., frame) as much as possible. In this case, we employ the GANs-style method to direct the sub-module to learn the *CFI*. In addition, all *CFI* vectors should be also irrelevant to all *SHI* and *SPI* vectors. The outputs of the CFID are three vectors, representing the *CFI* from the title, frame and motion features. In a similar vein, the loss function of the CFID is constructed as below:

$$\begin{aligned} X_{CFIG} &= G(Y_{CFI}) \\ \mathcal{L}_{cfG} &= -\mathbb{E}(D(X_{CFIG})) \\ \mathcal{L}_{cfD} &= \mathbb{E}(D(X_{CFIG})) - \mathbb{E}(D(X_{CFIR})) \\ \mathcal{L}_{cf} &= abs(\cos(X_{CFI}, Y)) \end{aligned} \tag{3}$$

where G and D are the generative and discriminative networks; X_{CFIG} is the generated *CFI* of modality X; X_{CFIR} is the original *CFI* of modality X; X_{CFI} and Y_{CFI} are the *CFI* of the modality X and its paired modality Y; $\mathcal{L}_{cfG}$ and $\mathcal{L}_{cfD}$ are the generative and discriminative network loss. Y is the set of all *SHI* and *SPI* vectors.

DEcision Module (DEM)

This module aims to predict ESV consumer response with content attributes, consisting of two sub-modules. They are the ConFlicted information Attention Module (CFAM) and Hierarchical Multitask Classification Module (HMCM). Specifically, we design the CFAM to determine how conflicted information of each pair of modalities (i.e., title vs. frame or title vs. motion) would be used. All information vectors given by the IDM will be taken into consideration when judging the conflicted information of which modality should be used or that of both modalities should be abandoned. For each pair of conflicted

information vectors, the CFAM will generate a vector (V) which contains only 0 or 1 and a binary weight (α), which is implemented by a Sigmoid function. The output for the CFAM is shown as below:

$$\begin{aligned} V &= F_1(T_{SHI}, T_{SPI}, T_{CFI}, m_{SHI}, m_{SPI}, m_{CFI}) \\ \alpha &= F_2(T_{SHI}, T_{SPI}, T_{CFI}, m_{SHI}, m_{SPI}, m_{CFI}) \\ Tm_{CFI} &= \alpha * [T_{CFI} * V + m_{CFI} * (1 - V)] \end{aligned} \tag{4}$$

where F_1 is the network to obtain the V; F_2 is the network to derive the α; T_{SHI}, T_{SPI}, T_{CFI} are the *SHI*, *SPI*, *CFI* of the ESV titles; m_{SHI}, m_{SPI}, m_{CFI} are the *SHI*, *SPI*, *CFI* of modality m, $m \in$ {motion, frame}; Tm_{CFI} is the output vector of the CFAM.

The outputs of the IDM and CFAM are fed into the HMCM. Since we examine the logically-dependent relations of content attributes and consumer viewing response, we develop a hierarchical multitask learning scheme. Following Chen et al. (2020), we consider the dependencies of feature representations and also the label dependencies. With the label transfer, the higher-level consumer response prediction task could utilize the labels of lower-level tasks (i.e., the content attributes classifications). In addition, we also find that the distributions of ESV content attributes and consumer response vary across different ESV categories (see Table 2). Thus, the 11 ESV categories are embedded to the vectors as an extra input information for the HMCM. The loss function of the HMCM is shown below, where $\mathcal{L}_{\text{n}}$ and $\mathcal{L}_{\text{resp}}$ are cross-entropy loss; n denotes one of the four ESV content attributes; *resp* denotes ESV consumer response.

$$\mathcal{L}_C = \mathcal{L}_n + \mathcal{L}_{resp} \tag{5}$$

Optimization

The information from the IDM is crucial for the DEM to make an accurate prediction. In this case, we will firstly train the IDM to extract the shared, special, and conflicted information. Secondly, given the structure of the CFAM and HMCM, training the DEM is a bilevel optimization problem (H. Liu et al. 2018). As such, we employ alternative training schemes to train the HMCM and CFAM one by one. In sum, we train our framework by three steps alternately for each epoch: (1) We train the IDM. (2) Then we train the IDM and DEM together but excluding the CFAM. (3) We train the CFAM only. Due to the GANs-style loss in the IDM for Step 1, there are two sub-processes to update all parameters: (1) fix all discriminator parameters and update other parameters, and (2) only update parameters in the discriminator.

The loss functions and the processing of parameters' update for Step 1 are as follows:

$$\begin{aligned} \mathcal{L}_{G,\min} &= \lambda_1 \mathcal{L}_{shG} + \lambda_2 \mathcal{L}_{sp} + \lambda_3 \mathcal{L}_{cfG} + \lambda_3 \mathcal{L}_{cf} \\ \mathcal{L}_{D,\min} &= \lambda_1 \mathcal{L}_{shD} + \lambda_3 \mathcal{L}_{cfD} \\ \theta_I, \theta_A &= \operatorname{argmin} \mathcal{L}_{G,\min}(\theta_I, \theta_A) \\ \theta_D &= \operatorname{argmin} \mathcal{L}_{D,\min}(\theta_D) \end{aligned} \tag{6}$$

where θ_I are parameters in the IDM other than those in the discriminator; θ_A are parameters in the FAM; θ_D are parameters of the discriminator in the IDM.

Similarly, for Step 2, there are two sub-processes:

$$\begin{aligned} \mathcal{L}_{G,\min} &= \mathcal{L}_C + \lambda_1 \mathcal{L}_{shG} + \lambda_2 \mathcal{L}_{sp} + \lambda_3 \mathcal{L}_{cfG} + \lambda_3 \mathcal{L}_{cf} \\ \mathcal{L}_{D,\min} &= \lambda_1 \mathcal{L}_{shD} + \lambda_3 \mathcal{L}_{cfD} \\ \theta_I, \theta_A, \theta_H &= \operatorname{argmin} \mathcal{L}_{G,\min}(\theta_I, \theta_A, \theta_H) \\ \theta_D &= \operatorname{argmin} \mathcal{L}_{D,\min}(\theta_D) \end{aligned} \tag{7}$$

where θ_H are parameters in the HMCM.

For Step 3, the parameters in the CFAM are updated as follows, together with the loss function:

$$\begin{aligned} \mathcal{L}_{\min} &= \mathcal{L}_C \\ \theta_{CA} &= \operatorname{argmin} \mathcal{L}_{\min}(\theta_{CA}) \end{aligned} \tag{8}$$

where θ_{CA} are parameters in the CFAM.

Experimental Setups

We evaluate our proposed framework using our ESVD. As mentioned before, we divide the 23,001 ESVs into three classes with short, medium, and long viewing duration. The number of ESVs in each class of viewing response is 2893, 11794 and 8314. For the four content attributes of ESVs, we use human-annotated labels as ground truth. We randomly split the whole dataset into training set (70%) and testing set (30%).

We extract modal features for all ESVs before training our framework. Frame features are generated by Inception-Resnet V2 with dimension size of 1536 (1 frame per second). For motion feature, multiple frames per 3.2 seconds are fed to S3D at once to extract dynamic features with dimension size of 512. For audio features, we use VGGish to learn acoustic deep features with 128-dimension size of every second. The BERT embedding technique is adopted to embed each word in video titles and category names into 768-dimensional vectors. Each LSTM network in the FAM has 2-layers with a hidden dimension size of 512.

The hyperparameters λ_1, λ_2, λ_3 are fine-tuned and set to 0.2, 0.2, 0.3 respectively. Our algorithm is optimized with the Root Mean Squared Propagation (RMSprop) algorithm for 150 epochs with a batch size of 512 and a learning rate of 0.0001, decaying ten times at every 30 epochs. To alleviate the class imbalance problem, we employ weighted cross-entropy loss. The weight of each label is inversely proportional to the number of samples of that label.

Following prior research on multi-label classification (Zhu et al. 2017), we report the macro-averaging F1 (F1-C, by calculating F1 for each class and then averaging per-class F1) and micro-averaging F1 (F1-O, by aggregating all classes and then computing the average F1) for prediction performance evaluation. For each ESV, the labels are predicted as positive if their confidence scores are greater than 0.5.

Experimental Results

We implement a series of experiments on ESVD to examine the effectiveness of our proposed framework.

First, we evaluate the effectiveness of the IDM and alignment methods. For this purpose, we only take the frame and title modality features as inputs to predict viewing responses to ESVs. We implement the baseline model as follows: (1) The frame and title features are pooled into a vector along sequential dimensions using the average pooling alignment (AvgPool), and then fed into its own fully connected layer (FC) (1-layer); (2) Two aligned feature vectors are sent to a 3-layer multilayer perceptron network (MLP); (3) The output vectors are concatenated followed by a 5-layer FC to make predictions. To examine our proposed IDM, the partial IDM (SHI + SPI + AvgPool) and the full IDM (SHI + SPI + CFI + AvgPool) are evaluated respectively. Further, we compare different alignment methods. As shown in Table 3, compared with the baseline, the full IDM increases the F1-C and F1-O by 3.8% and 3.6%, respectively. This is in line with our expectation since the IDM could obtain more informative high-level representations for predictions. We also show that using LSTM as the alignment method achieves the best prediction performance. The reason is that LSTM can fuse the sequential information more efficiently than AvgPool and NetVLAD (Arandjelovic et al. 2016). In the subsequent experiments, we consistently adopt the LSTM alignments.

Table 3. Evaluations of IDM and Alignment Methods on Viewing Response Predictions

IDM and Alignment Methods	*Macro F1-C*	*Micro F1-O*
Baseline + AvgPool	0.542	0.584
SHI + SPI + AvgPool	0.563	0.601
SHI + SPI + CFI + AvgPool	0.580	0.620
SHI + SPI + CFI + NetVLAD	0.556	0.586
SHI + SPI + CFI + LSTM	0.586	0.621

Second, we compare the effect of different modalities on ESV viewing response predictions using our framework. Table 4 shows that both motion and audio features improve the prediction performance. This is because they capture sellers' activity and acoustic information that the ESVs portray and communicate. Particularly, motion features improve 1.5% and 1.1% more than audio features do in terms of F1-C and F1-O, providing more complementary information. Using all the four modality features, we obtain the best result for ESV consumers' viewing response predictions.

Table 4. Effects of Different Modalities on Viewing Response Predictions		
Different Modalities	*Macro F1-C*	*Micro F1-O*
Frame + Title	0.586	0.621
Frame + Title + Motion	0.607	0.639
Frame + Title + Audio	0.592	0.628
Frame + Title + Motion + Audio	0.609	0.643

Third, we verify the effectiveness of the hierarchical multitask learning scheme in Table 5. Compared to the single task prediction, the multitask classification helps our framework gain better results due to the relations between ESV content attributes and consumer response. Moreover, with the logically-dependent relations between the content attributes and viewing response, the HMCM enables the low-level attributes prediction to directly influence the high-level viewing duration prediction to outperform the non-hierarchical multitask model by 1.3% and 0.9% in terms of F1-C and F1-O respectively. In addition, ESV category information could alleviate the class imbalance issue. After introducing category features, our proposed framework achieves the highest F1-C and F1-O for all labels and the difference between F1-C and F1-O of viewing duration predictions decreases from 0.032 to 0.025.

Table 5. Evaluations of DEM on Content Attributes and Viewing Response Predictions										
DEM	*ProdDesc*		*ProdDemo*		*Pleasure*		*Aesthetics*		*ViewDur*	
	F1-C	*F1-O*	*F1-C*	*F1-O*	*F1-C*	*F1-O*	*F1-C*	*F1-O*	*F1-C*	*F1-O*
IDiGM² (single task)	-	-	-	-	-	-	-	-	0.609	0.643
IDiGM² (multitask)	0.707	0.718	0.699	0.698	0.680	0.729	0.684	0.685	0.618	0.654
IDiGM² (HMCM w/o category)	0.702	0.716	0.700	0.704	0.685	0.729	0.681	0.682	0.631	0.663
IDiGM² (HMCM + category)	0.721	0.731	0.708	0.709	0.695	0.737	0.692	0.693	0.643	0.668

Conclusion

In this paper, we construct a large e-commerce short-video dataset with human-annotated content attributes and consumer response labels. Relying on the dataset, we reveal how ESV content attributes are related to consumers' viewing and purchase responses and show that the effects are contingent on different ESV types by purchase behaviors. Moreover, we propose a novel multimodal multitask framework to predict ESV viewing response. The effectiveness of our framework is evaluated by a series of experiments on our ESVD. The ability to understand and predict consumer response to ESVs is pivotal. Our work can help e-commerce as well as video sharing platforms to effectively manage and promote increasingly prevalent short-videos in many contexts. On future extensions, we would consider employing the feature maps from middle layers of each pre-trained model to obtain more useful information and to enhance our predictions.

References

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., and Sivic, J. 2016. "NetVLAD: CNN Architecture for Weakly Supervised Place Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307.

Bruce, N. I., Becker, M., and Reinartz, W. 2020. "Communicating Brands in Television Advertising," *Journal of Marketing Research* (57:2), pp. 236–256.

Chandrasekaran, D., Srinivasan, R., and Sihi, D. 2018. "Effects of Offline Ad Content on Online Brand Search: Insights from Super Bowl Advertising," *Journal of the Academy of Marketing Science* (46:3), pp. 403–430.

Chen, W., Tian, J., Xiao, L., He, H., and Jin, Y. 2020. "Exploring Logically Dependent Multi-Task Learning with Causal Inference," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2213–2225.

Chen, X., Chen, J., Ma, L., Yao, J., Liu, W., Luo, J., and Zhang, T. 2018. "Fine-Grained Video Attractiveness Prediction Using Multimodal Deep Learning on a Large Real-World Dataset," in *Companion Proceedings of the The Web Conference 2018 (WWW '18)*, pp. 671–678.

CNNIC. 2022. "49th Statistical Report on Internet Development," No. 49, China Internet Network Information Center, pp. 1–87.

Cohendet, R., Yadati, K., Duong, N. Q., and Demarty, C.-H. 2018. "Annotating, Understanding, and Predicting Long-Term Video Memorability," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pp. 178–186.

Demarty, C.-H., Sjöberg, M., Ionescu, B., Do, T.-T., Gygli, M., and Duong, N. 2017. "Mediaeval 2017 Predicting Media Interestingness Task," in *MediaEval Workshop*.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2018. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding," *ArXiv:1810.04805*. (https://arxiv.org/abs/1810.04805).

Elpers, J. L. C. M. W., Wedel, M., and Pieters, R. G. M. 2003. "Why Do Consumers Stop Viewing Television Commercials? Two Experiments on the Influence of Moment-to-Moment Entertainment and Information Value," *Journal of Marketing Research* (40:4), pp. 437–453.

Flavián, C., Gurrea, R., and Orús, C. 2017. "The Influence of Online Product Presentation Videos on Persuasion and Purchase Channel Preference: The Role of Imagery Fluency and Need for Touch," *Telematics and Informatics* (34:8), pp. 1544–1556.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. 2014. "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*.

Guitart, I. A., and Stremersch, S. 2021. "The Impact of Informational and Emotional Television Ad Content on Online Search and Sales," *Journal of Marketing Research* (58:2), SAGE Publications Sage CA: Los Angeles, CA, pp. 299–320.

Hagtvedt, H., and Patrick, V. M. 2008. "Art Infusion: The Influence of Visual Art on the Perception and Evaluation of Consumer Products," *Journal of Marketing Research* (45:3), pp. 379–389.

Hoyer, W. D., and Stokburger-Sauer, N. E. 2012. "The Role of Aesthetic Taste in Consumer Behavior," *Journal of the Academy of Marketing Science* (40:1), pp. 167–180.

Hussain, Z., Zhang, M., Zhang, X., Ye, K., Thomas, C., Agha, Z., Ong, N., and Kovashka, A. 2017. "Automatic Understanding of Image and Video Advertisements," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1705–1715.

Jiang, Y.-G., Wang, Y., Feng, R., Xue, X., Zheng, Y., and Yang, H. 2013. "Understanding and Predicting Interestingness of Videos," in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Jiang, Z. (Jack), and Benbasat, I. 2007. "The Effects of Presentation Formats and Task Complexity on Online Consumers' Product Understanding," *MIS Quarterly* (31:3), pp. 475–500.

Kuang, Q., Jin, X., Zhao, Q., and Zhou, B. 2020. "Deep Multimodality Learning for UAV Video Aesthetic Quality Assessment," *IEEE Transactions on Multimedia* (22:10), IEEE, pp. 2623–2634.

Kumar, A., and Tan, Y. (Ricky). 2015. "The Demand Effects of Joint Product Advertising in Online Videos," *Management Science* (61:8), pp. 1921–1937.

Lavie, T., and Tractinsky, N. 2004. "Assessing Dimensions of Perceived Visual Aesthetics of Web Sites," *International Journal of Human-Computer Studies* (60:3), Elsevier, pp. 269–298.

Lerche, L., Jannach, D., and Ludewig, M. 2016. "On the Value of Reminders within E-Commerce Recommendations," in *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, Halifax Nova Scotia Canada: ACM, July 13, pp. 27–35.

Liu, H., Simonyan, K., and Yang, Y. 2018. "Darts: Differentiable Architecture Search," *ArXiv:1806.09055*. (https://arxiv.org/abs/1806.09055).

Liu, X., Shi, S. W., Teixeira, T., and Wedel, M. 2018. "Video Content Marketing: The Making of Clips," *Journal of Marketing* (82:4), SAGE Publications Sage CA: Los Angeles, CA, pp. 86–101.

Liu, X., Zhang, B., Susarla, A., and Padman, R. 2020. "Go to YouTube and Call Me in the Morning: Use of Social Media for Chronic Conditions," *MIS Quarterly* (44:1b), pp. 257–283.

Liu, Y., Gu, Z., Ko, T. H., and Hua, K. A. 2018. "Learning Perceptual Embeddings with Two Related Tasks for Joint Predictions of Media Interestingness and Emotions," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pp. 420–427.

Ma, S., Zha, Z., and Wu, F. 2019. "Knowing User Better: Jointly Predicting Click-through and Playtime for Micro-Video," in *2019 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, pp. 472–477.

MacInnis, D. J., Rao, A. G., and Weiss, A. M. 2002. "Assessing When Increased Media Weight of Real-World Advertisements Helps Sales," *Journal of Marketing Research* (39:4), pp. 391–407.

MacKenzie, S. B., Lutz, R. J., and Belch, G. E. 1986. "The Role of Attitude toward the Ad as a Mediator of Advertising Effectiveness: A Test of Competing Explanations," *Journal of Marketing Research* (23:2), pp. 130–143.

Maclnnis, D. J., and Jaworski, B. J. 1989. "Information Processing from Advertisements: Toward an Integrative Framework," *Journal of Marketing* (53:4), pp. 1–23.

Miech, A., Alayrac, J.-B., Smaira, L., Laptev, I., Sivic, J., and Zisserman, A. 2020. "End-to-End Learning of Visual Representations from Uncurated Instructional Videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9879–9889.

Moe, W. W. 2003. "Buying, Searching, or Browsing: Differentiating Between Online Shoppers Using In-Store Navigational Clickstream," *Consumers in Cyberspace* (13:1), pp. 29–39.

Mowat, J. 2018. *Video Marketing Strategy: Harness the Power of Online Video to Drive Brand Growth*, Kogan Page Publishers.

Mulier, L. 2021. "This Way Up: The Effectiveness of Mobile Vertical Video Marketing," *Journal of Interactive Marketing* (55), pp. 1–15.

Munthiu, M.-C. 2009. "The Buying Decision Process and Types of Buying Decision Behaviour," *Sibiu Alma Mater University Journals. Series A. Economic Sciences* (2:4), pp. 27–33.

Olney, T. J., Holbrook, M. B., and Batra, R. 1991. "Consumer Responses to Advertising: The Effects of Ad Content, Emotions, and Attitude toward the Ad on Viewing Time," *Journal of Consumer Research* (17:4), pp. 440–453.

Park, M., Naaman, M., and Berger, J. 2016. "A Data-Driven Study of View Duration on Youtube," in *Proceedings of the International AAAI Conference on Web and Social Media* (Vol. 10), pp. 651–654.

Pavlou, Liang, and Xue. 2007. "Understanding and Mitigating Uncertainty in Online Exchange Relationships: A Principal-Agent Perspective," *MIS Quarterly* (31:1), p. 105.

Redi, M., O'Hare, N., Schifanella, R., Trevisiol, M., and Jaimes, A. 2014. "6 Seconds of Sound and Vision: Creativity in Micro-Videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4272–4279.

Song, D., Wang, S., and Ou, C. X. 2021. "How Do Video Features Matter in Visual Advertising? An Elaboration Likelihood Model Perspective," in *ICIS 2021 Proceedings*, pp. 1–16.

Southgate, D., Westoby, N., and Page, G. 2010. "Creative Determinants of Viral Video Viewing," *International Journal of Advertising* (29:3), pp. 349–368.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. 2016. "Rethinking the Inception Architecture for Computer Vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2818–2826.

Tan, Z., Zhang, S., Hong, N., Kuang, K., Yu, Y., Yu, J., Zhao, Z., Yang, H., Pan, S., Zhou, J., and Wu, F. 2022. "Uncovering Causal Effects of Online Short Videos on Consumer Behaviors," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, Virtual Event AZ USA: ACM, February 11, pp. 997–1006.

Teixeira, T., Wedel, M., and Pieters, R. 2012. "Emotion-Induced Engagement in Internet Video Advertisements," *Journal of Marketing Research* (49:2), pp. 144–159.

Tellis, G. J., MacInnis, D. J., Tirunillai, S., and Zhang, Y. 2019. "What Drives Virality (Sharing) of Online Digital Content? The Critical Role of Information, Emotion, and Brand Prominence," *Journal of Marketing* (83:4), SAGE Publications Sage CA: Los Angeles, CA, pp. 1–20.

Tong, L. C., Acikalin, M. Y., Genevsky, A., Shiv, B., and Knutson, B. 2020. "Brain Activity Forecasts Video Engagement in an Internet Attention Market," *Proceedings of the National Academy of Sciences* (117:12), pp. 6936–6941.

Vedula, N., Sun, W., Lee, H., Gupta, H., Ogihara, M., Johnson, J., Ren, G., and Parthasarathy, S. 2017. "Multimodal Content Analysis for Effective Advertisements on Youtube," in *2017 IEEE International Conference on Data Mining (ICDM)*, IEEE, pp. 1123–1128.

Vieira, V. A. 2010. "Visual Aesthetics in Store Environment and Its Moderating Role on Consumer Intention," *Journal of Consumer Behaviour* (9:5), pp. 364–380.

Wu, S., Rizoiu, M.-A., and Xie, L. 2018. "Beyond Views: Measuring and Predicting Engagement in Online Videos," in *Proceedings of the International AAAI Conference on Web and Social Media*, pp. 434–443.

Xie, J., Zhu, Y., Zhang, Z., Peng, J., Yi, J., Hu, Y., Liu, H., and Chen, Z. 2020. "A Multimodal Variational Encoder-Decoder Framework for Micro-Video Popularity Prediction," in *Proceedings of The Web Conference 2020*, Taipei Taiwan: ACM, April 20, pp. 2542–2548.

Yang, J., Zhang, J., and Zhang, Y. 2021. "First Law of Motion: Influencer Video Advertising on TikTok," *SSRN Working Paper*.

Ye, K., Buettner, K., and Kovashka, A. 2018. "Story Understanding in Video Advertisements," *ArXiv:1807.11122*.

Zhou, M., Chen, G. H., Ferreira, P., and Smith, M. D. 2021. "Consumer Behavior in the Online Classroom: Using Video Analytics and Machine Learning to Understand the Consumption of Video Courseware," *Journal of Marketing Research* (58:6), pp. 1079–1100.

Zhu, F., Li, H., Ouyang, W., Yu, N., and Wang, X. 2017. "Learning Spatial Regularization with Image-Level Supervisions for Multi-Label Image Classification," *ArXiv:1702.05891*.