First Law of Motion:
Influencer Video Advertising on TikTok †

Jeremy Yang, Juanjuan Zhang, Yuhan Zhang

March 28, 2021

Abstract

This paper develops an algorithm to predict the effect of influencer video advertising on product sales. We propose the concept of motion-score, or m-score, a summary statistic that captures the extent to which a product is advertised in the most engaging parts of a video. We locate product placement with an object detection algorithm, and estimate pixel-level engagement as a saliency map by fine-tuning a deep 3D convolutional neural network on video-level engagement data. M-score is then defined as the pixel-level, engagement-weighted advertising intensity of a video. We construct and evaluate the algorithm with around 40,000 influencer video ads on TikTok, the largest short video platform of the world. We leverage variation in video posting time to identify the causal effect of video ads on product sales. Videos of higher m-score indeed lift more sales. This effect is sizable, robust, and more pronounced among impulsive, hedonic, or inexpensive products. We trace the mechanism to influencers’ incentives to promote themselves rather than the product. We discuss how various stakeholders in entertainment commerce can use m-score in a scalable way to optimize content, align incentives, and improve efficiency.

†Authors are listed alphabetically. Jeremy Yang (jeryang@hbs.edu) is an Assistant Professor of Business Administration at the Harvard Business School. Juanjuan Zhang (jjzhang@mit.edu) is the John D. C. Little Professor of Marketing at the MIT Sloan School of Management. Yuhan Zhang (yhz_zhang@hotmail.com) is a Lecturer at the Beijing Technology and Business University.
Every body perseveres in its state of being at rest ... except insofar as it is compelled to change its state by forces impressed.  

Sir Isaac Newton  

*The Mathematical Principles of Natural Philosophy* (1687)

1 Introduction

“The next Amazon competitor is going to look like a social or video app, not a shopping app,” says Connie Chan, general partner of venture capital firm Andreessen Horowitz.\(^1\) E-commerce is moving beyond utilitarian and search-driven platforms to embrace more entertaining and discovery-driven platforms. On the latter type of platforms, the lines between content and commerce are blurry. Content creators, or influencers, often engage with users and sell to them at the same time. This mixing of entertainment and commerce has been described as “entertainment commerce.” One might even say that the letter *e* is now standing for *entertainment* in this emerging form of e-commerce.\(^2\)

TikTok is one of the major platforms leading this transformation. Its core feature of short-video sharing has attracted a massive following. As the most downloaded app of the world in 2020, TikTok has over 1.4 billion active users around the globe, and has collected over $16 billion in advertising revenues in 2020.\(^3\) E-commerce integration is prominent on TikTok, especially in its original Chinese version.\(^4\) Product sellers routinely pay influencers to place products directly in their videos, and users are able to make a purchase without leaving the app. In the latest news, Walmart’s pilot test of shoppable TikTok videos, in which influencers feature their favorite Walmart products, is seen as the

---

\(^1\)[https://twitter.com/conniechan/status/1266476997699493889.](https://twitter.com/conniechan/status/1266476997699493889.)

\(^2\)This integration of entertainment and commerce is happening in both directions. On one hand, social media platforms such as Facebook, Instagram, and YouTube are introducing shoppable content. On the other, e-commerce platforms such as Alibaba, Amazon, and Walmart are adding entertainment features.


\(^4\)Also known as Douyin, the Chinese version of TikTok has over 600 million active users (Reuters Business News, September 15, 2020). See Kaye et al. (2020) for a detailed comparison of Douyin and the international version of TikTok.
retail giant’s ambitious move into the video-powered, entertainment commerce space.⁵

Despite its sharp rise, how influencer video advertising contributes to product sales is unclear. There is not yet a systematic way to predict an influencer video ad’s sales lift, meaning the incremental sales conversion attributed to the ad.⁶ As a result, product sellers have relied on influencer engagement metrics (such as the number of likes, shares, and comments) for campaign management. On TikTok, many product sellers would simply choose an engaging influencer, and leave it to the discretion of the influencer to design a video ad. The result has been less than ideal. Anecdotal evidence abounds where an influencer video ad is highly engaging, but does a poor job of lifting sales.⁷

The goal of this paper is to develop a method to predict the effect of influencer video ads on product sales. Drawing on the theory of bottom-up attention in cognitive psychology and neuroscience, we propose an algorithm for computing a compact, intuitive, and interpretable summary statistic of video ads, one that turns unstructured data into structured information, is predictive of sales lift, and is measurable prior to the ad’s release. We call this summary statistic motion-score, or m-score for brevity.

The motion concept comes from the following analogy. Newton’s first law of motion states that motion is possible only when a force is applied onto an object (assuming the object is at rest). Both the force and the object are necessary for motion to occur. In a video ad, we hypothesize that both content engagement and product placement are necessary for sales lift to materialize. More specifically, variation in engagement over pixel space and time of a video creates a force field with varying strength, and that the overall motion (sales lift) is strong if the object (product placement) appears where the force (content engagement) is strong. By borrowing these concepts from one of the most fundamental principles in physics, we aspire to establish a basic principle for influencer video advertising: place the product in the most engaging space and time of a video, or

---

⁶The company that provided us data emphasizes strong industry demand for such predictive tools.
make the video more engaging where and when the product appears.

Technically, we define a video ad’s m-score as its average pixel-level engagement-weighted advertising intensity. We compute m-score in three steps.

First, we construct a three-dimensional (3D) matrix that we call the *content engagement heatmap*. The three dimensions are the height and width of each video frame in pixels, and the length of the video in seconds. A video’s content engagement heatmap is a pixel-level saliency map, which outputs the gradient of video-level engagement with respect to each pixel in the video. We estimate the content engagement heatmap by fine-tuning a deep 3D convolutional neural network on video-level engagement data.

In the second step, we construct a 3D *product placement heatmap*, which shows whether the product being advertised is present at a given pixel in a given frame of the video. We estimate the product placement heatmap by matching an image of the product to each frame of the video with an object detection algorithm called the “scale-invariant feature transform.”

In the third step, we compute m-score as the Frobenius inner product of the two 3D matrices, normalized by the total number of pixels of the video. This normalized inner product can be interpreted as the average pixel-level engagement-weighted advertising intensity, or in other words, the extent to which the product is shown in more spatiotemporally engaging parts of a video ad.

We hypothesize that a video ad with higher m-score is going to be more effective in lifting sales. It is important to note that m-score measures the interaction, or more specifically, the complementarity between content engagement and product placement. A video that is engaging or features the product throughout does not necessarily have a high m-score.

We test our approach by analyzing a proprietary dataset that contains around 40,000 influencer video ads on the original Chinese version of TikTok (referred to as TikTok for brevity hereafter) and their corresponding product sales revenue on Taobao from March
through June 2019. Indeed, the data reveal no correlation between video engagement metrics and product sales. This echoes the industry’s criticism of engagement as an inadequate predictor of sales conversion in entertainment commerce.

To better measure sales conversion, we estimate the causal effect of influencer video ads on product sales via the staggered difference-in-differences method, leveraging the variation in video posting time for identification. Consistent with our hypothesis, m-score positively moderates the sales lift of a video ad. A one standard deviation increase in m-score is associated with a 12% increase in sales revenue of the advertised product. Notably, engagement or advertising intensity (as measured by the total number of pixels in which the product appears in a video) alone has no moderation effect on sales lift. These results highlight the unique predictive power of the m-score metric – making the video more engaging or increasing product placement unilaterally does not help; it is the complementarity between the two that drives sales.

Our findings are robust with respect to alternative ways to construct the algorithm and to identify sales lift. To better understand the applicability of our algorithm, we conducted a supplementary survey to classify advertised products in our data. Our algorithm is more effective in product categories associated with impulse purchases, hedonic consumption, or lower prices. These products are also popular choices for entertainment commerce, given its focus on facilitating unplanned product discovery.

Last but not least, we explore the behavioral mechanism behind our algorithm. What drives m-score? One explanation is the agency problem inherent to entertainment commerce. Influencers are incentivized to promote themselves rather than the product. As such, they may not want to allocate the most engaging space and time of their videos to the advertised product. To test this mechanism, we collected yet another auxiliary dataset, in which influencers advertise their own products. Consistent with the agency-problem interpretation, m-score tends to be higher in these videos than in videos where

---

8Owned by Alibaba, Taobao is the world’s largest e-commerce website, and the major platform on which products advertised in TikTok influencer videos were sold during the time of our data.
influencers advertise for another party, particularly when it comes to impulsive and hedonic products.

Our algorithm can be practically valuable in several ways. First, we invested heavily in algorithm calibration, such that m-score can be easily computed for a video ad in future applications. The algorithm is also highly scalable; its only data requirement is the video ad itself once the algorithm is constructed. This means influencers can use the algorithm as an automated tool to test and improve their videos in the creative process for better sales lift later on. Second, the content engagement heatmap developed in this paper can itself serve as a useful tool to improve video engagement. Third, m-score introduces a new contractual instrument to the entertainment commerce space. Product sellers can use m-score to screen candidate videos or directly write a contract based on it. Influencers can use m-score to signal their business involvement beyond what engagement metrics are able to communicate. Platforms can design a myriad of ways to use m-score for more accurate attribution and more efficient allocation. After all, the m-score concept is built upon the two pillars of entertainment commerce – entertainment, and commerce.

2 Related Research in Marketing

Our paper is inspired by, and contributes to, several streams of marketing research. First, we address a problem in influencer marketing (Avery and Israeli 2020). Influencer marketing is a $10 billion industry in 2020 with a whopping 50% growth rate. It is a marketing strategy that uses the influence of key individuals, or opinion leaders, to drive consumers' brand awareness and purchase decisions (Brown and Hayes 2008). Social media is the main channel through which influencers influence. A social media influencer is first a content creator then a marketer; she/he produces valuable content to captivate and cultivate a sizable number of followers, and monetizes their attention.

---

Previous work on influencer marketing has studied the effect of influencer attributes on self-reported purchase intent. For example, Lou and Yuan (2019) found that influencer’s trustworthiness, attractiveness, and similarity to the followers affect followers’ brand awareness and purchase intent. Schouten et al. (2020) showed that influencer endorsement is more effective than celebrity endorsement, and that the effect is mediated by higher perceived similarity and trust.

We contribute by studying one of the latest forms of influencer marketing – influencer video advertising. Despite its rapid growth, influencer video advertising has been understudied in academia. One exception is Rajaram and Manchanda (2020), who analyzed the relationship between YouTube influencer video ad content and video views, interaction rates, and sentiment. In comparison, we focus on sales conversion – we develop an algorithm to predict sales lift from influencer video ad content.

Our focus on sales adds to the discussion of engagement versus conversion. Teixeira et al. (2014) found that ad entertainment, as measured by viewers’ facial reactions, increases viewing but has an inverted U-shaped effect on purchase intent. Tucker (2015) showed with a survey experiment that purchase intent is 10% lower for each million views of a video ad, an effect likely driven by outrageous video ads purposely designed to command attention and provoke sharing. John et al. (2017) found in both field and lab experiments that liking a brand on Facebook has no positive impact on consumer attitudes or purchases; what matters is consumers’ preexisting preference for the brand. Using actual sales data, we also find that engagement does not guarantee conversion. Furthermore, we propose and validate a novel process that connects engagement with conversion. We find that what affects sales is not the level of engagement per se, but whether the advertised product is featured in the more engaging parts of a video.
Our paper is also related to the marketing literature on video content design. One prominent stream of research focuses on measuring moment-to-moment (MTM) viewer behaviors and relating them to various outcomes of video content. Many innovative MTM measurement strategies have been developed, including handheld devices (Polsfuss and Hess 1991), “feeling monitor” computers (Baumgartner et al. 1997), eye tracking (Wedel and Pieters 2008, Teixeira et al. 2010), electroencephalography (Barnett and Cerf 2017), facial expression tracking (Liu et al. 2018), functional Magnetic Resonance Imaging (Tong et al. 2020), and viewer live comments analysis (Zhang et al. 2020).

We contribute to this video-content-design literature along four dimensions. First, the literature has focused on movies or standard video ads. We study a new type of content – video ads produced by influencers. Influencer video ads can be fundamentally different from traditional video content. Indeed, we find that influencers’ incentive to promote themselves rather than the product does affect content design. Second, many methods proposed in this literature require collection of MTM data for new videos in order to forecast their market outcomes. We use historical observational data on video-level engagement to infer pixel-level engagement without directly measuring them. This means our algorithm is scalable and can be applied directly to new videos. Third, much of the literature has focused on time-series data to capture the temporal dimension of video features. We made a nontrivial investment to extend the analysis to the pixel-moment level. This most granular, spatiotemporal approach to video content design helps reveal new insights. Fourth, the literature typically used pre-defined features to represent video content, whereas we take a data-driven approach without relying on hand-crafted features.
– and we do so without sacrificing the interpretability of our algorithm. We turn to the algorithm, its theoretical motivation, and its construction in the next section.

3 Algorithm Construction

The m-score concept is motivated by the distinctive shopping process on entertainment commerce platforms. Users typically come to these platforms for entertainment. On TikTok, for instance, users often passively browse a stream of video feeds without a clear goal of searching or purchasing a product. However, purchase interest can be activated in the process of consuming a video ad, if the advertised product happens to grab user attention.

This shopping process is closely related to theories of bottom-up attention in cognitive psychology and neuroscience (e.g., Milosavljevic and Cerf 2008). Bottom-up attention is a rapid, automatic form of selective attention that depends on the intrinsic properties of the input. It is also known as saliency-based attention, indicating that the more salient an object, the higher the probability of it being noticed. In contrast, top-down attention is a volitional, focal, task-dependent mechanism, often compared to a spotlight that enhances processing of the selected item (Koch 2004). As such, top-down attention tends to characterize the buying process in traditional search-driven product markets, whereas bottom-up attention more likely pertains to entertainment commerce platforms.

Drawing upon the theories of bottom-up attention, our hypothesis behind m-score is that, other things being equal, the more salient and engaging an advertised product is in an influencer video ad, the more effective the video ad will be in lifting sales. To operationalize this idea, we propose a three-step algorithm:

---

12 See discussions at, for example, https://www.oberlo.com/blog/tiktok-ads.

13 The seminal paper of Mitchell and Olson (1981) found that consumers’ attitude towards an ad can mediate their attitude towards the advertised brand. Our hypothesis complements their theory; we argue that attitude towards the ad, as measured by engagement, has a greater influence on attitude towards the brand if the brand is advertised in a more salient and engaging way.
1. Compute a pixel-level content engagement heatmap over the video ad to identify the most salient spots of the video.

2. Compute a pixel-level product placement heatmap over the video ad to identify when and where the product is featured in the video.

3. Compute m-score as the normalized inner product of the two heatmaps to capture the average engagement-weighted advertising intensity of the video.

We explain these three steps in detail in the following sections.

### 3.1 Content Engagement Heatmap

We first estimate a content engagement heatmap for each video ad, which is a 3D matrix that captures the spatiotemporal variation of content engagement in the video. The three dimensions of the content engagement heatmap are the height and width of each video frame in pixels, and the length of the video in seconds. Specifically, we train a deep 3D convolutional neural network (CNN) on historical video-level engagement data, and extract a saliency map over the input video.

The CNN architecture is suitable for our problem because it is well-known to be particularly good at image understanding (see Malik and Singh 2019 for a tutorial). We take a transfer learning approach by first extracting features from video frames with a CNN pre-trained on ImageNet (namely, MobileNetV2, Sandler et al. 2018), then feeding the feature sequence into a 3D convolution layer which accounts for the temporal dependencies across frames (e.g., Tran et al. 2015). This approach is also known as fine-tuning.

We take the transfer learning approach for two reasons. First, the pre-trained network is highly optimized for its performance on image recognition, which is directly relevant to our task. Transferring the knowledge encoded in this pre-trained network to our con-

---

14ImageNet (http://www.image-net.org) is a database of images with 1,000 class labels. It is considered as the industry standard for training and testing image classification algorithms. MobileNetV2 is a computationally light yet effective network, with a top-1 accuracy of 0.713 and top-5 accuracy of 0.901.
text is thus more computationally efficient. Second, building on a pre-trained network reduces the number of parameters to be estimated and mitigates the risk of overfitting. Nevertheless, because the bottom layers of a CNN tend to detect more general features of an image, whereas the top layers are more task-specific, we make the last two layers of the pre-trained network trainable to better tailor to our application.

For the main analysis, we use each video’s number of shares as the measure of engagement. There is a common sentiment that shares are stronger signals of engagement than likes and comments, as users are willing to endorse shared videos in their social networks. However, our results are robust if we use likes and comments to measure engagement (see Section 6.1). We also partial out influencer and product fixed effects and audio features from video-level engagement prior to training. This allows us to focus on the effect of influencer video ad content itself.

We train the CNN on 30,000 videos, validate it on 2,000 videos, and test it on 5,000 holdout videos (we offer more details of the data in Section 4). To appreciate the magnitude of the raw data for pixel-level analysis, consider a typical TikTok video. It is most commonly 15 seconds in length. Each second consists of up to 60 frames. Each frame of standard resolution on TikTok contains 1,080 * 1,920 pixels. Finally, each pixel has 3 RGB (Red, Green, and Blue) color channels. As a result, one typical TikTok video would contain a total of 15s * 60fps * 1,080p * 1,920p * 3 = 5,598,720,000 pixel values. We thus make several data simplifications to speed up the training process. We only use the first 15 seconds of a video, sub-sample videos to one frame per second, and resize each frame to a dimension of 224p * 224p. This allows each video to be represented as a much more feasible (15, 224, 224, 3) numerical array.

In the end, our full CNN has over 2 million variables (with each pixel value at a given

---

15For example, https://socialmediaweek.org/blog/2017/10/social-media-metrics-compared-valuable.
16MobileNetV2 is optimized on images of such dimension, which is also the standard image size for many widely used computer vision algorithms. We resize our video frames accordingly to make sure MobileNetV2 has the best performance, besides speeding up training. However, our algorithm should accommodate any image size in principle.
color channel being an independent variable), more than 4 million trainable parameters, and takes in more than 60 billion data points.\textsuperscript{17} We train the CNN with graphics processing units (GPUs) on a high-performance computing (HPC) cluster. See Online Appendix A for more details on the network structure and the training process.

After training the CNN, we use it to extract saliency maps on videos held out for downstream analysis (videos in the “sales panel”; see Section 4). A saliency map (Simonyan et al. 2013) is a heatmap over an original image that represents the gradient of the outcome variable with respect to this image. The value at each pixel on the saliency map corresponds to the partial derivative of the outcome variable with respect to that pixel. In our case, the outcome variable is video-level engagement, and the magnitude of the derivatives tells us how much video-level engagement changes with respect to changes in pixels of the input image, or equivalently, which pixels need to be changed to affect video-level engagement. We interpret a high absolute value of the derivative as high content engagement at that particular pixel. We ignore the sign and focus on the magnitude because increasing or decreasing the value of a pixel simply means changing its intensity along a particular color channel, which does not have an intrinsic, directional meaning.\textsuperscript{18} See Online Appendix B for more details on the saliency map.

3.2 Product Placement Heatmap

In the second step, we estimate a product placement heatmap, which is a 3D matrix of the same dimension as the content engagement heatmap. We do so by matching an advertised product’s image to each frame of a video to estimate when and where the product is placed. We use the scale-invariant feature transform (SIFT) algorithm (Lowe 1999) for product detection.

SIFT is a popular algorithm for object detection, matching features across different

\textsuperscript{17}There is a paper famously titled “I Just Ran Four Million Regressions” (Sala-i Martin 1997). Here we just ran one regression with four million parameters.

\textsuperscript{18}See http://www.cknuckles.com/rgbsliders.html for an interactive example.
images to identify the presence of an object from a cluttered scene. The key challenge is to make sure the key features of an object are robust to changes in scale, rotation, illumination, and viewpoint. The solution is intuitive. First, the “essence,” called keypoints, of both the reference image (product) and target image (video frame) are extracted; these keypoints are invariant to rotation and re-scaling of the image. Then the keypoints are matched between the reference and target images based on the distance of their characteristics, called keypoint descriptors.

Usually, a ratio test is performed on each matched keypoint to access its quality. The idea is the following. For a given keypoint, multiple matches with different distances can be found. One way to determine if the best match (the one with the shortest distance) is a good match is by looking at how it compares with the second-best match. If the two are too similar, then the best match is more likely to be noise. If the two are different enough, this means the best match is more likely to be distinct and thus a good match. Following convention, we use 0.75 as the cutoff, and consider the product to be present at a given pixel if the ratio between the best match and the second-best match is below this cutoff.\textsuperscript{19} The resulting product placement heatmap is a 3D matrix of binary values, where 1 indicates product presence at a pixel and 0 indicates absence. See Online Appendix C for more details on the product placement heatmap.

### 3.3 Computing M-Score

In the third step, we combine the content engagement heatmap and the product placement heatmap to calculate m-score. Let $e_{hwsv}$ be the (continuous) pixel-level value in the 3D content engagement heatmap and $p_{hwsv}$ be the (binary) pixel-level value in the 3D product placement heatmap. The symbols $h$, $w$, $s$, and $v$ denote height, width, time (in seconds), and video, respectively. We define a video’s m-score as the normalized inner

\textsuperscript{19}\url{https://docs.opencv.org/master/dc/dc3/tutorial/pymatcher.html}
product of the two maps:

\[ m_v := \frac{1}{HWS} \sum_{h,w,s} e_{hwsv} \cdot p_{hwsv}, \]  

(1)

where \( H, W, \) and \( S \) are the total height, width, and length of a video in seconds, respectively. Their product, \( HWS \), is thus the total number of pixels, or the volume, of a video. As discussed, we interpret m-score as the average engagement-weighted advertising intensity of a video; the inner product captures the complementarity between content engagement and product placement. We summarize the algorithm in Figure 1.

Figure 1. Summary of the Algorithm

Two remarks on our algorithm are in order. First, we train a 3D CNN on video content data using video-level engagement as the outcome variable. The number of parameters to estimate far exceeds the training sample size. This is a common feature of deep learning models and do not necessarily imply overfitting (e.g., Zhang et al. 2017). Nevertheless, we take several actions to mitigate overfitting concerns. We (1) construct the algorithm using a large sample of videos, (2) use transfer learning and downsample the video data to reduce the number of parameters to estimate, (3) use effective regularization methods.
such as dropout (Srivastava et al. 2014; see Online Appendix A), and (4) check for overfitting on the validation and test samples. Reassuringly, as we discuss in Online Appendix A, validation and testing results suggest no signs of overfitting.

Second, the ultimate goal of the paper is to predict product sales lift from influencer video ad content. The question is why we do not directly train a 3D CNN on video content data using product sales as the outcome variable. We can, but choose not to. This is because CNN results are typically difficult to interpret (Zhang and Zhu 2018). We instead compute m-score as an interim summary statistic that is succinct, behaviorally meaningful, and interpretable. Theoretically, this allows us to achieve greater clarity on what lifts sales. Practically, knowing what lifts sales sheds light on the famously challenging problem of marketing attribution (e.g., Testwuide 2020).

In what follows, we test our algorithm with data. Specifically, we test whether influencer video ads of high m-score lift more sales, other things being equal. We present the data in the next section.

4 Data

We test our algorithm using data we collect from TikTok. Launched in 2016 by Chinese tech company ByteDance, TikTok is the world’s leading user-generated short video platform. With over 2.6 billion app installs so far, TikTok is the most downloaded app of the world in 2020, and is the most liked among the top ten iOS apps in the United States during 2020. By January 2021, TikTok has over 1 billion monthly active users. Each day, an average TikTok user would open the app eight times and spend 52 minutes on the platform, browsing content that largely center around entertainment.20

We focus on the Chinese version of TikTok because of its mature ecosystem around influencer video advertising. There is an established marketplace called Xingtu, where

product sellers contract with influencers to advertise their products. By January 2021, Xingtu has attracted over 330,000 influencers and over 760,000 product sellers.\textsuperscript{21}

Two distinctive features characterize this marketplace. First, engagement is the centerpiece of the ecosystem. It determines how influencers price their video ads, how product sellers search for influencers, and how product sellers monitor ad performance. Second, influencers have significant discretion in designing their video ad content. In a typical ad creation process, an influencer drafts an ad script, makes the video upon seller confirmation of the script, and posts the ad upon seller confirmation of the video. Product sellers are thus able to influence ad content to some degree. However, there are many video design aspects that are controlled by the influencer. As such, there is no clear way for product sellers to predict sales lift from an ad. They pay for engagement, in the (sometimes shattered) hopes that engaging influencers would lift sales.

We test whether our algorithm can help predict sales lift. To do so, we need data on influencer video ads (and ideally many of them for algorithm construction), on video engagement metrics, and on actual sales of the advertised products. We are fortunate to have developed such a dataset via collaboration with an entertainment commerce company. Notably, in current industry practice, content and engagement data are usually stored in one system (i.e., social media platforms), while sales data are typically stored in another (e.g., e-commerce websites). It is valuable to be able to connect these data sources.\textsuperscript{22}

Specifically, our dataset is a matched sample from two separate sources: (1) a video dataset that contains TikTok influencer video ads, their corresponding engagement metrics measured at a daily level, as well as influencer characteristics, (2) a product dataset that contains product images, product category label, price, discount, and sales revenue (summed over the previous 30 days measured at a daily level) on Taobao. TikTok and

\footnotesize{\textsuperscript{21}https://star.toutiao.com.  
\textsuperscript{22}For instance, Lee et al. (2018) studied advertising content and engagement on Facebook, and mentioned not having access to sales data as a limitation.}
Taobao are the main advertising and sales channels for product sellers during the time of our sample. This is yet another desirable feature of our data which helps us attribute product sales lift on Taobao to video ads on TikTok.

Data from these two sources from March through June 2019 are cleaned and matched. Some products are missing category information. We use product title and the non-missing category labels to train a machine learning model that predicts the missing product categories. The model has an accuracy of 82% in the test sample (see Online Appendix D for details).

The resulting dataset contains 42,856 video ads. Among them, 5,856 video ads have matching product sales data. Among these products, 2,734 have only one video ad. We will focus on these product-video pairs in subsequent analysis for clean attribution.\(^\text{23}\) We call the sales panel dataset of these 2,734 product-video pairs the sales panel, which we will hold out and use to test our algorithm’s ability to predict sales lift.\(^\text{24}\) As discussed, the remaining 37,000 video ads are used to construct the algorithm, with 30,000 videos randomly assigned to the training set for the CNN, 2,000 videos randomly assigned to the validation set, and the remaining 5,000 to the holdout test set. This leaves us with a high-power test of our algorithm – we construct it from one subsample of videos, and test its predictive power on a different subsample, which helps examine the external validity of the algorithm.\(^\text{25}\)

Table 1 presents summary statistics of video engagement metrics: the number of likes, comments, and shares.\(^\text{26}\) We do observe video engagement at the daily level. However, engagement takes time to grow. To capture each video’s ultimate level of engagement,

\(^{23}\)If a product has multiple video ads, it is nontrivial how to attribute sales lift to each ad. See Du et al. (2019) for a model of “multi touch attribution.”

\(^{24}\)For each product in the panel, sales is observed at the daily level although there are missing observations for technical reasons that are believed to have occurred randomly.

\(^{25}\)Although our sample is not a random subset of all video ads on TikTok, the data collection and matching process is orthogonal to video and product popularity. No video or product is included or excluded systematically based on their engagement or sales data. Therefore, our findings can plausibly generalize.

\(^{26}\)We also observe the number of plays for each video. However, play volume can be a noisy measure of engagement because it does not capture how much time users actually spend on a video. Nevertheless, we control for play volume in subsequent analysis.
ment, we use its last observed value in our data. These engagement metrics are statistically indistinguishable between videos in the sales panel and videos for algorithm construction, except that the former received more likes.

Table 1. Summary Statistics of Video Engagement Data

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Videos in the Sales Panel</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># Likes</td>
<td>2,734</td>
<td>38,413</td>
<td>111,408</td>
<td>0</td>
<td>3,614.5</td>
<td>1,831,709</td>
</tr>
<tr>
<td># Comments</td>
<td>2,734</td>
<td>538</td>
<td>2,039</td>
<td>0</td>
<td>84</td>
<td>71,068</td>
</tr>
<tr>
<td># Shares</td>
<td>2,734</td>
<td>926</td>
<td>4,965</td>
<td>0</td>
<td>80</td>
<td>166,821</td>
</tr>
<tr>
<td>Videos for Algorithm Construction (Training, Validation, and Test Sets)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likes</td>
<td>37,000</td>
<td>30,745</td>
<td>104,351</td>
<td>0</td>
<td>2,516.5</td>
<td>2,553,627</td>
</tr>
<tr>
<td>Comments</td>
<td>37,000</td>
<td>519</td>
<td>2,236</td>
<td>0</td>
<td>62</td>
<td>71,697</td>
</tr>
<tr>
<td>Shares</td>
<td>37,000</td>
<td>1,077</td>
<td>6,082</td>
<td>0</td>
<td>80</td>
<td>195,563</td>
</tr>
</tbody>
</table>

Note: Each engagement metric is at the video level.

Table 2 presents summary statistics of sales revenue, as well as prices and discounts, of all products in the sales panel. Average 30-day sales revenue of products in our data is 234,688 RMB, or 33,964 USD based on the average 2019 currency exchange rate of 6.91:1. There is again wide variation in revenue, price, and discount amount across products.

Table 2. Summary Statistics of Product Sales Data (Sales Panel)

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Sales Revenue</td>
<td>2,734</td>
<td>234,688</td>
<td>3,419,870</td>
<td>0</td>
<td>8,163</td>
<td>29,166,580</td>
</tr>
<tr>
<td>Price</td>
<td>2,734</td>
<td>1,069</td>
<td>38,867</td>
<td>0.01</td>
<td>68</td>
<td>2,019,515</td>
</tr>
<tr>
<td>Discount</td>
<td>2,734</td>
<td>102</td>
<td>507</td>
<td>0</td>
<td>20</td>
<td>13,901</td>
</tr>
</tbody>
</table>

Note: Each variable is at the product level, and measured in RMB. A product’s sales revenue is its revenue summed over the previous 30 days observed at the daily level. A product’s average sales revenue is its sales revenue averaged over its observed days in our sample. There is no variation in price or discount over the duration of our sample at the product level.

We present further details of the sales panel in Online Appendix E. In summary, engagement and sales show sizable variation across videos (Figure E.1). Most influencers post one video ad, although there is a wide distribution (Figure E.2). This variation allows us to control for influencer fixed effects in subsequent analysis. Nevertheless, we report influencer summary statistics in Table E.1. In addition, the most common video length
is 15 seconds (Figure E.3), whereas video posting date is widely distributed (Figure E.4), a fact we will leverage for causal identification of sales lift. The most common category in the data is food, followed by makeup; there is a range of prices although the average price for most categories falls below 300 RMB (Figure E.5).

One pattern in the data that is particularly revealing is the lack of significant correlation between video engagement and product sales. Figure E.6 presents scatter plots of the relationship between residualized engagement metrics (the number of likes, comments, and shares) and residualized product sales, after controlling for price, discount, video play volume, influencer fixed effects, and product category fixed effects. None of the engagement measures is significantly associated with sales. This result suggests that using video engagement to choose influencers and evaluate advertising effectiveness can be misleading. Our algorithm addresses this problem. We present its test and results in the next section.

5 Algorithm Evaluation – Main Results

In this section, we first present the computational results of our algorithm. We then proceed to the main test, of whether influencer video ads of higher m-score lift more sales. We further explore whether the effect is stronger in some categories than others. Last, we investigate the behavioral mechanism that drives m-score.

5.1 Computational Results of the Algorithm

For each video ad, the algorithm outputs a 3D content engagement heatmap, a 3D product placement heatmap, and an m-score. Table 3 presents the video-level summary statistics of these three output metrics for videos in the sales panel, which is the sample we will use to evaluate the algorithm. In the table, “computed engagement,” termed to differentiate it from actual engagement, is a video’s sum of pixel-level engagement
scores, and product placement is a video’s sum of pixels in which the product appears. To facilitate interpretation, we normalize all three output metrics to the interval of [0, 1] in this table and in subsequent analysis. Notably, a good fraction of videos do not have a high m-score.

Table 3. Summary Statistics of Video-Level Computed Engagement, Product Placement, and M-Score

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Computed Engagement</td>
<td>2,734</td>
<td>0.41</td>
<td>0.12</td>
<td>0</td>
<td>0.40</td>
<td>1</td>
</tr>
<tr>
<td>Product Placement</td>
<td>2,734</td>
<td>0.11</td>
<td>0.04</td>
<td>0</td>
<td>0.11</td>
<td>1</td>
</tr>
<tr>
<td>M-Score</td>
<td>2,734</td>
<td>0.08</td>
<td>0.03</td>
<td>0</td>
<td>0.08</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: The sample consists of all videos in the sales panel. All three output metrics are at the video level and normalized to [0,1]. Video-level computed engagement and product placement scores are aggregated from pixel-level scores.

To further visualize these computational results, we present average pixel-level engagement and product placement values within a video frame (Figures 2a and 2b) and over the duration of a video (Figures 2d and 2e). For completeness, we analogously present pixel-level m-score, computed as the pixel-level product of engagement and product placement, averaged over space and time (Figures 2c and 2f).

We see a pattern. The most engaging region of a video frame tends to be its center, with the left side of the frame being slightly more engaging than the right, possibly because of users’ tendency to read from left to right in this market (see also Spalek and Hammad 2005). Similarly, on average, products tend to appear in the center of the frame, and pixel-level m-score also tends to peak around the center. However, we cannot simply conclude that we should put the product in the center. The most engaging regions vary from frame to frame. In fact, the pixel-level correlations between frames are low; even between adjacent frames where transition tends to be more gradual, pixel-level correlations are 0.076, 0.053, and 0.052 for engagement, product placement, and m-score, respectively. This suggests that there is nontrivial incremental information in our 3D heatmaps compared with their average, 2D representations.
Over the duration of a video, engagement tends to start low in the first two seconds, peak in the middle, and decline sharply in the last three seconds. Products tend to appear the most in the beginning and the second half of a video. M-score rises and falls but sharply drops near the end of the video, possibly due to fading engagement in these moments. However, we again cannot simply conclude that products should be placed in moments where average engagement peaks. These dynamics vary significantly across videos. The gray regions in the figures represent values within 0.1 standard deviation from the mean, which span a wide range already. These observations again highlight the incremental value of our algorithm, which captures rich heterogeneity across space, time, and videos. We formally test our algorithm in the following section.
5.2 Influencer Video Ads of Higher M-Score Lift More Sales

We test our main hypothesis in two steps. We first quantify the causal impact of influencer video ads on sales. We then examine whether this causal sales lift is stronger among influencer video ads of higher m-score.

We identify the causal impact of influencer video ads on sales using the staggered difference-in-differences approach (e.g., Stevenson and Wolfers 2006, Liu et al. 2019). We estimate the following specification:

\[
\text{Sales}_{vid} = \alpha \cdot \text{Post}_{vid} + \text{Video}_v + \text{Influencer}_i + \text{Day}_d + \epsilon_{vid}. \tag{2}
\]

The dependent variable, \(\text{Sales}_{vid}\), is the prior-30-day sales revenue of a product, which has a video ad \(v\) posted by influencer \(i\), as measured on day \(d\).\(^{27}\) As discussed, we focus on products that have only one video ad, so that \(v\) also indexes the product. Meanwhile, recall that influencers vary in the number of videos they post, and so we index videos and influencers separately.

The treatment variable is \(\text{Post}_{vid}\), which equals 1 if video ad \(v\) by influencer \(i\) is posted by day \(d\), and 0 otherwise. The parameter \(\alpha\) measures the average sales revenue lifted by an influencer video ad. Leveraging the panel structure of the data, we include video/product fixed effects \(\text{Video}_v\) and influencer fixed effects \(\text{Influencer}_i\) to control for unobserved heterogeneity across videos/products and influencers, respectively. We also include day fixed effects \(\text{Day}_d\) to capture common time trends in sales. As mentioned, the rich variation in video posting date allows us to separately identify the treatment effect of video ads from such time trends.\(^{28}\) Lastly, \(\epsilon_{vid}\) is the error term.

Column (1) of Table 4 presents the ordinary least squares (OLS) regression result. Overall, posting an influencer video ad does not significantly lift product sales. This is

\(^{27}\)Our results are robust if we use log-transformation of sales as the dependent variable.

\(^{28}\)There could be a reverse-causality concern if influencers post a video ad in anticipation of product sales lift. This concern is less likely to apply in our setting because, as discussed, influencers are mainly motivated by engagement metrics instead of product sales.
an alarming result, given that sellers pay nontrivial amounts to advertise their products (influencers on average charge 19,530 RMB per video; see Table E.1). This result further highlights the importance of being able to predict sales lift before investing in an influencer video ad.

Table 4. Effect of Influencer Video Ads on Sales

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post</td>
<td>-14.51</td>
<td>-133.90**</td>
<td>-14.71</td>
<td>-21.52</td>
<td>-113.64</td>
</tr>
<tr>
<td>(15.04)</td>
<td>(41.51)</td>
<td>(15.26)</td>
<td>(17.15)</td>
<td>(89.01)</td>
<td></td>
</tr>
<tr>
<td>Post × M-Score</td>
<td>1,525.41**</td>
<td>(494.36)</td>
<td>1,420.65**</td>
<td>(503.16)</td>
<td></td>
</tr>
<tr>
<td>Post × Engagement</td>
<td>1.37E-4</td>
<td>(1.78E-3)</td>
<td>-6.21E-4</td>
<td>(2.08E-3)</td>
<td></td>
</tr>
<tr>
<td>Post × Product Placement</td>
<td></td>
<td>140.21</td>
<td>132.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(164.48)</td>
<td>(164.32)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post × Covariates</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Video, Influencer, Day FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>173,515</td>
<td>173,515</td>
<td>173,515</td>
<td>173,515</td>
<td>173,515</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Note: The sample is the sales panel, where each product has one video ad. Each observation is at the product-day level. OLS. Dependent variable is a product’s prior-30-day sales revenue in 1,000 RMB. Post is the treatment dummy variable of whether an influencer video ad is posted. All columns control for video/product, influencer, and day fixed effects (FE). Column (5) also controls for the interaction between Post and covariates including product characteristics (price and discount) and influencer characteristics (see Table E.1). *p < 0.05, **p < 0.01, ***p < 0.001.

In the second step to test our main hypothesis, we examine heterogeneous treatment effects, of whether m-score positively moderates the effect of influencer video ads on sales. We estimate the following specification:

\[
Sales_{vid} = \alpha \cdot Post_{vid} + \beta \cdot Post_{vid} \times M-Score_v + Video_v + Influencer_i + Day_d + \epsilon_{vid},
\]

where \(M-Score_v\) is the m-score of video \(v\). Note that the main effect of \(M-Score_v\) cannot be separately identified from video fixed effects. Our main hypothesis is equivalent to the coefficient \(\beta\) being positive. As column (2) of Table 4 shows, this is indeed the case. Influencer video ads with higher m-score are significantly more effective in lifting sales.
To test whether it is the complementarity between content engagement and product placement that predicts sales lift, we also check whether engagement alone or product placement alone would have a similar moderation effect. As column (3) of Table 4 shows, more engaging video ads (as measured by the actual number of shares) are not more effective in lifting sales. This reaffirms industry observations and our correlation analysis findings, that video engagement fails to predict sales lift. Column (4) of Table 4 shows that simply advertising the product more intensively in the video does not help either. This result echoes the “advertising avoidance” literature, where viewers dislike excessive ads when their viewing purpose is entertainment (e.g., Elpers et al. 2003, Wilbur 2016). In a sense, our algorithm suggests one possible way to advertise sparingly but wisely.

Last, we include m-score, engagement, and product placement simultaneously as moderating variables for column (5) of Table 4. We also control for the interaction between Post and a rich set of covariates, including product characteristics (price and discount) and influencer characteristics (see Table E.1). We call this comprehensive model the “main specification.” We reach qualitatively the same conclusions. The moderation effect of m-score is highly significant and sizable. To put the effect magnitude in context, one standard deviation increase in m-score is on average associated with a 28,413 RMB (about 4,111 USD) increase in monthly sales revenue of a product. This is about a 12% increase of sales revenue for products in our sample.

5.3 M-Score Predicts Better for Impulsive, Hedonic, and Inexpensive Products

We have shown that, consistent with our hypothesis, m-score predicts sales lift. In this section, we investigate whether m-score predicts better in some product categories than others. This will help us better understand where and why our algorithm works.

To characterize the 11 product categories (except “other”) in our sample, we ran-
domly surveyed 175 TikTok users on Wenjuanxing, a major survey platform in China. We ask these users to rate each product category along three dimensions: product purchase being impulsive vs. deliberate, products being utilitarian vs. hedonic, and advertising informativeness, all on the scale of 1-7. We also collected users’ demographic information including gender, age, education, and income. The distributions of these demographics are similar between our survey sample and users on TikTok.

We summarize the average ratings across product categories in Figure F.1 of the Online Appendix. The results appear intuitive. For example, buying in the furniture category needs the most deliberation, products in the entertainment category are the most hedonic, and ads in the electronics category are the most informative. More impulsive categories also tend to be more hedonic in our data. This is not surprising given that hedonic goals are known to drive impulsive behaviors (e.g., Ramanathan and Menon 2006).

To test whether the predictive power of m-score varies systematically by category, we perform median splits of the sales panel based on category ratings on each of the three dimensions, as well as category price level which we observe from our data. We then re-estimate our main specification on the split subsamples. The estimation results reveal a noticeable pattern. M-score’s ability to predict sales lift is significant only if the product category is more impulsive, more hedonic, or less expensive (Table 5), and the conclusion is robust with respect to where the equal sign goes in the median split. The result for informativeness depends on where the equal sign goes, suggesting that sample size may be a main driver of statistical significance.

These results are consistent with the behavioral process underlying the m-score concept: users come to the platform for entertainment; an engaging ad may activate their bottom-up attention, which translates into purchase if buying does not require much cognitive processing. In other words, a higher m-score is more likely to mean higher sales.

---

29These are users who answered yes to the following question: have you watched any ads on TikTok?
30http://www.show5jia.com/weixin/20190304/5c7c897b92902.pdf.
31The conclusion is also robust if we perform median split at the product level, which mitigates the concern that statistical significance is an artifact of sample size.
Table 5. Predictive Power of M-Score by Product Category

<table>
<thead>
<tr>
<th></th>
<th>Deliberate ≤ Median</th>
<th>Deliberate &gt; Median</th>
<th>Hedonic ≤ Median</th>
<th>Hedonic &gt; Median</th>
<th>Price ≤ Median</th>
<th>Price &gt; Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post</td>
<td>-119.87</td>
<td>80.78</td>
<td>82.70</td>
<td>-126.27</td>
<td>-138.50</td>
<td>-38.80</td>
</tr>
<tr>
<td>Post × M-Score</td>
<td>2,171.40***</td>
<td>-700.27</td>
<td>-767.42</td>
<td>2,175.66***</td>
<td>3,425.63***</td>
<td>-552.76</td>
</tr>
<tr>
<td>Post × Engagement</td>
<td>-3.49E-4</td>
<td>3.21E-3</td>
<td>3.72E-3</td>
<td>-3.47E-4</td>
<td>-1.62E-3</td>
<td>-4.71E-4</td>
</tr>
<tr>
<td>Post × Product Placement</td>
<td>42.39</td>
<td>-32.35</td>
<td>-12.67</td>
<td>164.08</td>
<td>-35.17</td>
<td>314.59</td>
</tr>
<tr>
<td>Post × Covariates</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Video, Influencer, Day FE</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>143,151</td>
<td>30,364</td>
<td>31,209</td>
<td>142,306</td>
<td>90,859</td>
<td>82,656</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.98</td>
<td>0.71</td>
<td>0.72</td>
<td>0.98</td>
<td>0.78</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Note: The entire sample is the sales panel, where each product has one video ad. Each observation is at the product-day level. OLS. Dependent variable is a product’s prior-30-day sales revenue in 1,000 RMB. Post is the treatment dummy variable of whether an influencer video ad is posted. All columns control for video/product, influencer, and day fixed effects (FE), and the interaction between Post and covariates including product characteristics (price and discount) and influencer characteristics (see Table E.1). *p < 0.05, **p < 0.01, ***p < 0.001.

lift if users’ buying decision is more automatic, intuitive, and less deliberate (Kahneman 2011). This type of buying decision in turn is more likely to happen in impulsive, hedonic, or inexpensive product categories. These categories commonly appear on entertainment commerce platforms; they together represent over 90% of products in our TikTok data (see Figure E.5 of the Online Appendix). Notably, video as a popular content format of entertainment commerce may be more effective at influencing “low involvement” users compared with print media (Liu et al. 2020). As such, we expect m-score to be a particularly suitable predictor of sales conversion in entertainment commerce.

5.4 Exploring the Mechanism behind M-Score

We have seen that m-score varies across videos and this variation matters in predicting sales lift. The next natural question is what drives the variation in m-score. We test two explanations: influencers’ ability to produce a video ad of high m-score, and their willingness to do so.

For the ability explanation, one possibility is that influencers with less experience...
do not know how to make an effective video ad. We test this explanation by regressing m-score on measures of influencer experience, including the number of video ads the influencer has posted, and the number of days since the influencer’s first post, controlling for other influencer characteristics (see Table E.1).\textsuperscript{32} We find no statistically significant association between m-score and these influencer experience measures, contrary to the ability explanation.

The willingness explanation is more subtle, yet reflects a unique feature of influencer video ads in entertainment commerce. Given the centrality of engagement metrics in this ecosystem and given users’ general distaste for seeing ads during entertainment (e.g., Elpers et al. 2003, Wilbur 2016), influencers may not have every incentive to advertise the product. In light of the m-score concept, this means influencers may avoid placing the product in the most engaging spots of the video, worrying that doing so might hurt their follower engagement.

To test this explanation, we collected a separate sample of 77 video ads where influencers advertise their own products.\textsuperscript{33} Following the willingness explanation, better incentive alignment should mean higher m-score for these video ads. We thus regress m-score on an indicator variable of whether the influencer is advertising her/his own product, while controlling for product price, discount, and influencer characteristics. As shown in Table 6, we find that m-score is significantly higher in video ads for influencers’ own products than others’ products in impulsive and hedonic categories. In other words, incentive misalignment between the influencer and the seller is more of a problem in categories where m-score matters more. This result is intuitively appealing because, if m-score does not matter, the influencer might as well focus on being engaging. The ex-

\textsuperscript{32}There are many missing values in the number of days since the influencer’s first post. To conserve sample size, unlike other influencer characteristics in Table E.1, we do not include this characteristic as a moderator variable of advertising effect.

\textsuperscript{33}Analogously, Levitt and Syverson (2008) test incentive misalignment in the housing market by comparing home sales whereby agents sell for others vs. themselves. See Wernerfelt et al. (2021) for a formal treatment of advertising agency issues. See also Pei and Mayzlin (2019) for a theory in which influencers receive financial incentives to review products.
ception is that m-score is not significantly higher in influencers’ own ads for cheaper products. A possible explanation is that lower prices diminish the monetary return from product sales, which shifts influencers’ focus to engagement even if they are advertising their own products.

Table 6. M-Score by Influencer Advertising Own versus Others’ Products

<table>
<thead>
<tr>
<th></th>
<th>Deliberate ≤ Median</th>
<th>Deliberate &gt; Median</th>
<th>Hedonic ≤ Median</th>
<th>Hedonic &gt; Median</th>
<th>Price ≤ Median</th>
<th>Price &gt; Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advertises Own Product</td>
<td>7.32* (3.18)</td>
<td>−7.45 (6.53)</td>
<td>−9.31 (7.38)</td>
<td>10.02* (4.59)</td>
<td>8.04 (7.27)</td>
<td>1.10 (4.97)</td>
</tr>
<tr>
<td>Covariates</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>1,726</td>
<td>1,085</td>
<td>1,173</td>
<td>1,638</td>
<td>1,449</td>
<td>1,362</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Note: The sample consists of all videos in the sales panel and the videos in which influencers advertise their own products. Each observation is at the video level. OLS. Dependent variable is m-score times 1,000 for easier presentation of results. Controlling for product price and discount and influencer characteristics (see Table E.1). *p < 0.05, **p < 0.01, ***p < 0.001.

Taken together, the evidence seems to be more consistent with the willingness explanation of incentive misalignment between the influencer and the product seller. Our algorithm can provide a solution to this problem by quantifying to what extent the influencer is effectively advertising the product. Admittedly, influencers who advertise for others vs. for themselves may be inherently different. As such, the mechanism analysis presented in this section is exploratory. A comprehensive study of the mechanism will be a worthy topic for future research.

6 Robustness Checks and Extensions

In the previous section, we validated our algorithm, examined its applicability across product categories, and explored its underlying mechanism. In this section, we extend the analysis to check the robustness of our algorithm. We check two aspects of robustness, with respect to the construction of the algorithm, and with respect to the causal identification of sales lift.
6.1 Alternative Measurement of Content Engagement

In the main analysis, we use the number of shares as the outcome variable to train the 3D CNN and to extract saliency maps. A first robustness check is to reconstruct the algorithm using the number of likes and comments instead. We re-estimate the main specification based on these alternative measures of m-score. Columns (1) and (2) of Table 7 present the results. The moderation effect of m-score remains positive and significant.

Table 7. Alternative Measurement of Content Engagement

<table>
<thead>
<tr>
<th></th>
<th>Sales Revenue</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td>Post × M-Score (# Likes)</td>
<td>269.58**</td>
</tr>
<tr>
<td></td>
<td>(100.72)</td>
</tr>
<tr>
<td>Post × M-Score (# Comments)</td>
<td>116.42*</td>
</tr>
<tr>
<td></td>
<td>(48.36)</td>
</tr>
<tr>
<td>Post × M-Score (Unsupervised)</td>
<td>590.11***</td>
</tr>
<tr>
<td></td>
<td>(141.80)</td>
</tr>
<tr>
<td>Post × Engagement</td>
<td>-6.83E-4</td>
</tr>
<tr>
<td></td>
<td>(2.11E-3)</td>
</tr>
<tr>
<td>Post × Product Placement</td>
<td>294.42</td>
</tr>
<tr>
<td></td>
<td>(156.91)</td>
</tr>
<tr>
<td>Post × Covariates</td>
<td>Yes</td>
</tr>
<tr>
<td>Video, Influencer, Day FE</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>173,515</td>
</tr>
<tr>
<td>Adjusted R²</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Note: See Table 4 note.

So far, we have followed the “supervised” approach to construct the content engagement heatmap, using video content as input and video-level engagement (shares, likes, or comments) as output. Pixel-level engagement is thus determined in a data-driven way; a pixel will have high engagement if a small change in its value affects video-level engagement by a large amount. We next check if the algorithm is robust if we estimate content engagement using the “unsupervised” approach.

The motivation of the unsupervised approach is that content engagement may be correlated with the intrinsic properties of the images themselves. The more salient regions in an image may disproportionately affect overall engagement. Past research has
also shown that saliency measures based on intrinsic properties of images predict actual gaze and eye movement (e.g., Itti 2005, Dupont et al. 2016). In addition, the unsupervised approach does not rely on video engagement data; pixel-level engagement is determined entirely by the image itself. This is a helpful feature that expands the applicability of our algorithm to environments where historical engagement data are unavailable (e.g., new entertainment commerce platforms). Therefore, as a robustness test, we use the intrinsic properties of the images (the statistically distinct areas of an image, such as high contrast locations and edges of objects) as a proxy for pixel-level engagement.\footnote{We use the algorithm in the saliency module in OpenCV to estimate the unsupervised content engagement heatmap. See Figure B.1b of the Online Appendix for an illustration and Hou and Zhang (2007) for more details of the algorithm.} Reassuringly, as column (3) of Table 7 shows, the m-scores metric based on unsupervised learning continues to predict higher sales lift.

6.2 Human Face Predicts Engagement

The saliency map outputs the more engaging regions of a video in a bottom-up, data-driven way but does not necessarily offer an explanation. Substantively, one might wonder what content features drive engagement. Past research has identified human face as an engaging object that attracts likes and comments on social media (Bakhshi et al. 2014, Li and Xie 2020). Therefore, as a simple sanity check of our algorithm, we identify human faces in the videos to see if they are predictive of pixel-level engagement.

We use a face detection algorithm to locate human faces across all frames in a video.\footnote{https://pypi.org/project/face-recognition.} For each frame, the algorithm outputs the location of boxes that contain a human face. Similar to the product placement heatmap, we estimate a face placement heatmap where the values inside the boxes are coded as 1 and values outside are coded as 0. We then compute the correlation between 3D pixel-level content engagement values with the indicator variable of whether a face is present in a pixel. The result indeed shows a positive and
significant correlation \((r = 0.02, p < 0.001)\). This adds face validity (pun intended) to our content engagement heatmap because we are now more confident that it is uncovering the more engaging parts of a video as we intended.

### 6.3 Pre-Trends and Dynamic Treatment Effects

In this and the following sections, we test the robustness of our causal identification of sales lift. We first test for parallel pre-trends between treated and not-yet-treated products, an assumption that underlies the staggered difference-in-differences approach. We also investigate dynamic treatment effects to allow advertising effects to vary over time. To do so, for each product we include 30-day leads and lags of the treatment of ad posting. Using \(t_{vi}\) to denote the date when influencer \(i\) posts a video ad \(v\), and using \(1\) to denote the indicator function, we estimate the standard dynamic specification:

\[
Sales_{vid} = \sum_{e=-30}^{30} \alpha_e \cdot 1\{d - t_{vi} = e\} + Video_v + Influencer_i + Day_d + \epsilon_{vid}.
\]  

Figure 3 presents the estimated coefficients \(\alpha_e\) for \(e \in \{-30, \cdots, 30\}\). None of these coefficients are significantly different from zero. This result supports the parallel-trends assumption. Meanwhile, it reinforces our earlier finding that positing video ads on average does not significantly lift sales, even if we allow the advertising effect to vary over time. This result also mitigates the serial correlation concern of difference-in-differences analysis (Bertrand et al. 2004).

### 6.4 Alternative Specification of Difference-in-Differences

A recent stream of methodological papers on difference-in-differences argued against using the two-way fixed effects (TWFE) specification in settings with more than two time periods and variation in treatment timing (e.g., Goodman-Bacon 2018, Callaway and Sant’Anna 2020, Sun and Abraham 2020). The criticism is that TWFE does not recover
easy-to-interpret causal parameters. In general, a TWFE regression recovers a weighted average of some underlying treatment effect parameters that vary by group and time, but some of the weights on these parameters can be negative (known as the “negative weight problem”). This can cause issues such as the treatment effect being positive for all units with TWFE giving a negative result. Even in cases where the weights are not negative, they are driven by the TWFE estimation strategy and are sensitive to the size of each group, the timing of treatment, and the total number of time periods.

Callaway and Sant’Anna (2020) proposed a solution. The first step is to estimate group-time average treatment effects, where a group is defined by the time period when cross-sectional units are first treated. Then these group-time average treatment effects can be aggregated to a single parameter with weights that are application-dependent. In our application, we are primarily interested in aggregating all group-time average treatment effects into an overall treatment effect and see if it differs by m-score (see Online Appendix G for more details).
We follow this procedure and estimate the overall treatment effect separately for the full sample and the subsample of products with m-score in the top 10% range. As Table 8 shows, the results are consistent with our main result: the treatment effect of posting a video ad is significantly positive for videos with high m-score but insignificant on the full sample of videos.

### Table 8. Sales Lift by M-Score

<table>
<thead>
<tr>
<th>Sales Revenue</th>
<th>All Video Ads</th>
<th>High-M-Score Ads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Post</td>
<td>28.43</td>
<td>472.70**</td>
</tr>
<tr>
<td></td>
<td>(56.41)</td>
<td>(152.02)</td>
</tr>
<tr>
<td>Covariates</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Observations</td>
<td>173,515</td>
<td>16,981</td>
</tr>
</tbody>
</table>

Note: Callaway and Sant’Anna (2020) specification. High m-score means in the top 10% of the m-score distribution. Controlling for product characteristics (price and discount) and influencer characteristics (see Table E.1). *p < 0.05, **p < 0.01, ***p < 0.001.

### 6.5 Randomization Inference

To further substantiate our causal identification strategy, we use Fisher’s randomization inference as a placebo test. The idea is the following. If the treatment effect is truly present, when we take random draws of video posting time and re-estimate the model, most of the time we should not see an interaction effect between treatment and m-score that is as sizable as the one estimated with the actual data.

We make 1,000 such bootstrap draws of video posting time, and re-estimate the moderation effect of m-score. As shown in Figure 4, these estimated effects are mostly below the effect from the main analysis (the dotted vertical line) and the difference is significant ($p = 0.010$); they are also not significantly different from zero ($p = 0.315$). This result gives us further confidence that the moderation effect of m-score we have uncovered is not due to chance, and indeed captures the predictive power of our algorithm.
Figure 4. Randomization Inference for the Moderation Effect of M-Score

Note: The figure presents the distribution of the coefficients of Post \( \times \) M-score (in 1,000 RMB) over 1,000 permutations of the video ads’ posting time. The dashed vertical line is the coefficient estimated from the observed data.

7 Concluding Remarks

In this paper, we propose an algorithm to compute a summary statistic called m-score, which (1) turns unstructured video data into structured information, and (2) predicts the effect of influencer video ads on product sales. This summary statistic computes the pixel-level engagement-weighted advertising intensity of a video ad, with an intuitive interpretation – it captures to what extent a product is shown in the most engaging part of the video, or how engaging the video is when and where the product is shown.

We construct and evaluate the algorithm using a proprietary dataset of around 40,000 TikTok influencer video ads and their corresponding product sales. The data suggest that video engagement is not predictive of product sales. This reaffirms the industry’s concern that engagement is an inadequate success metric for influencer advertising. We then connect engagement with product placement via m-score, and show that video ads with higher m-score are more effective in lifting sales. The effect is significant and sizable. One standard deviation increase in m-score is associated with about $4,111 increase in
monthly sales. This is about a 12% increase of sales in our sample.

In further analysis, we find that m-score predicts especially well for impulsive, hedonic, and inexpensive products. As expected, these products are also commonly featured on TikTok and entertainment platforms in general. We show that our findings are robust to different ways to construct the algorithm and different causal identification strategies. We also show evidence that the agency problem, but not influencer experience, may explain the variation in m-score, particularly for impulsive and hedonic product categories.

The m-score metric can potentially unlock a myriad of practical applications. A key practical advantage of m-score is that it can be computed before a video ad is released, without relying on in-consumption user data such as eye tracking or live comments. This means that the algorithm is highly scalable, and can be used to evaluate a large number of candidate videos very quickly.

More specifically, influencers can use m-score to aid video content development in real time. Product sellers can use m-score as a novel contractual instrument. For example, product sellers can compensate influencers based on the m-score of their video ads. In comparison, the current industry practice of engagement-based compensation has been shown to be ineffective, whereas sales-based compensation makes influencers accountable for product sales but exposes them to various factors beyond their control (such as perceived product quality, which is difficult to contract on). In this sense, m-score can serve as a metric to help clarify the attribution of sales outcome between product sellers and influencers. Finally, entertainment commerce platforms can leverage m-score to launch various features to improve transaction efficiency. For example, a platform can highlight m-score as a key performance index of influencers. Providing m-score alongside engagement metrics can help product sellers choose influencers with richer information.

There are several directions for future research. First, as discussed, it will be interesting to study various applications of the algorithm and track their impact on the entertainment commerce industry. Second, m-score is learned mainly through the visual com-
ponent of a video ad while controlling for the audio features. How to better integrate the two in an interpretable way is an open question. Lastly, it will be meaningful to explore the generalizability of the algorithm. We validated the algorithm in the context of influencer video ads, where m-score fundamentally matters because it captures bottom-up attention in entertainment commerce, and because influencers may strategically choose m-score as a result of incentive misalignment. However, the general principle of making product placement engaging should extend to other forms of video advertising as well. It will be encouraging if the algorithm is able to predict sales lift based on the mere content of a generic video ad.
References


Tkachenko, Y. and K. Jedidi (2020). What personal information can a consumer facial image reveal? Implications for marketing ROI and consumer privacy. *SSRN 3616470*.


A Details on the 3D Convolutional Neural Network (3D CNN)

We use a deep 3D CNN and gradient-based saliency map to estimate the content engagement heatmap from observed video-level engagement data (number of shares, likes, or comments). We use the number of shares as the outcome variable in the main analysis, and verify robustness using the numbers of likes and comments. As discussed in the paper, each video in our data is represented as a (15, 224, 224, 3) numerical array, where 15 is the length of the video in seconds, (224, 224) is the height and width of each video frame in pixels, and 3 is the number of RGB color channels. The output is a single numerical value representing the predicted number of shares of the video. This is a supervised learning problem.

The key building blocks of a CNN are convolution layers. A convolution layer uses filters with weights that are trainable to transform the input images by representing them in a more abstract feature space that captures more general properties of the images (e.g., the presence of an edge or face). What properties are captured depends on what the network is trained for. Multiple convolution layers can be stacked on top of each other, interspersed with other non-trainable layers such as max pooling layers (to reduce the dimension of feature space), non-linear activation layers (to perform a simple non-linear transformation of input values), and dropout layers (to randomly set some weights to zero to avoid overfitting). After many layers of transformation, the feature maps are flattened into a vector and fed into a fully connected layer for the final classification or regression task.

CNNs have been used to analyze images for marketing research in a few recent papers (e.g., Hartmann et al. 2019, Zhang and Luo 2019, Liu et al. 2020, Tkachenko and Jedidi 2020, Zhang et al. 2020). These papers are built upon 2D CNNs. We refer interested readers to “A Comprehensive Guide to Convolutional Neural Networks – The ELI5 Way”
(Saha 2018) for a visual introduction that animates what each layer does.¹

In our paper, we use a 3D CNN to account for the additional time dimension of video content. We highlight the difference between a standard 2D convolution and a 3D convolution in Figure A.1. In a 2D convolution, the filter always has the same depth as the input \( L \); it only slides across the spatial dimensions of the input \( H \) and \( W \), which means the output is a 2D matrix. In contrast, the filter in 3D convolution has a variable depth \( d < L \), and in addition to sliding across the spatial dimensions, it also slides across the depth dimension outputting a 3D matrix.

Figure A.1. 2D versus 3D Convolution (Tran et al. 2015)

In Figure A.2 below, we illustrate the architecture of our 3D CNN use case, where the interim layers are adapted from the 2D CNN illustration of Saha (2018).

More specifically, we use MobileNetV2 (Sandler et al. 2018) pre-trained on ImageNet to extract features from each frame (in a time distributed manner), and stack a 3D con-

---

volution layer with 128 units on top of the extracted feature sequence to account for the
temporal dependency across frames. We also include a max pooling layer to reduce the
dimension of the feature space, and a dropout layer which has been shown to be partic-
ularly effective at reducing overfitting (Srivastava et al. 2014). Because the top layers of
MobileNetV2 are trained for image recognition purposes, which is relevant but not the
same as engagement prediction, we also allow the weights in the top two layers to be
updated in the training process, whereas weights in other layers are kept frozen. The
architecture of our network on top of MobileNetV2 is summarized in Figure A.3.²

Figure A.3. CNN Layers on Top of MobileNetV2

<table>
<thead>
<tr>
<th>Layer (type)</th>
<th>Output Shape</th>
<th>Param #</th>
</tr>
</thead>
<tbody>
<tr>
<td>time_distributed_2</td>
<td>(None, 15, 7, 7, 1280)</td>
<td>2257984</td>
</tr>
<tr>
<td>conv3d_2 (Conv3D)</td>
<td>(None, 13, 5, 5, 128)</td>
<td>4423808</td>
</tr>
<tr>
<td>max_pooling3d_2 (MaxPooling3) (None, 6, 2, 2, 128)</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>dropout_2 (Dropout)</td>
<td>(None, 6, 2, 2, 128)</td>
<td>0</td>
</tr>
<tr>
<td>flatten_2 (Flatten)</td>
<td>(None, 3072)</td>
<td>0</td>
</tr>
<tr>
<td>dense_2 (Dense)</td>
<td>(None, 1)</td>
<td>3073</td>
</tr>
</tbody>
</table>

Total params: 6,684,865
Trainable params: 4,429,441
Non-trainable params: 2,255,424

Our final CNN has more than 4 million trainable parameters. We train it on 30,000
videos and validate it on 2,000 videos, with GPUs on a high performance computing
cluster using TensorFlow (Keras).³ Product/video and influencer fixed effects, and audio
features are projected out of the engagement metrics before training.⁴ Figure A.4 sum-
marizes the training and validation loss statistics. The mean absolute percentage error
(MAPE) on the holdout test set of 5,000 videos is 23.31%.

²MobileNetV2 has 155 layers, hence our full network has 155 + 6 = 161 layers. See Sandler et al. (2018)
for more details on the architecture of MobileNetV2.
³https://keras.io.
⁴We extract a numerical representation (amplitude) of the sound wave in each video. The raw sampling
rate is 44,100 per second. We down-sample it to 100 evenly spaced observations per audio file.
Figure A.4. Training and Validation Loss

Note: Statistics is over 20 epochs. The y-axis is mean squared error in $10^6$. The training losses are higher than validation losses likely because a dropout layer is used in training but not validation.
B Details on the Saliency Map

A saliency map is a gradient-based visualization method for CNNs (Simonyan et al. 2013). It takes a trained network and computes the gradient of the outcome with respect to a given input image. Each entry of the map represents the partial derivative of the outcome with respect to that particular pixel in the input image. Usually, the absolute value of the gradient is used on the map. A high absolute value suggests that a small change in that pixel will lead to a big change in the outcome. For color images, there are three channels: red, green and blue, or RGB. It is typical to compute the gradient for each channel, and take the maximum across channels as the final value for that pixel. The eventual output of a saliency map is of the same dimension as the input image, except that the three color channels, as explained, are flattened into one layer.

We adapt the saliency map to videos, which are sequences of images (frames). Importantly, instead of computing the gradient with respect to pixels frame by frame, we do so with respect to pixels in the whole video. This allows us to capture any dependency across video frames when deciding which pixels are driving the outcome. More formally, we define pixel-level content engagement as:

$$e_{hws} := \left| \frac{\partial \hat{f}}{\partial x_{hws}} \right|$$

where $\hat{f}$ is the trained 3D CNN, and $x_{hws}$ is the pixel value at location $(h, w, s)$ in a video, with $h$ being the index for height (in pixels), $w$ for width (in pixels), and $s$ for time (in seconds).

We use a saliency map to compute pixel-level content engagement and call it the (supervised) content engagement heatmap. It is supervised because the saliency map builds on a 3D CNN trained on video-level engagement data. In Section 6 of the main text, we also discuss an unsupervised approach to content engagement heatmap that only requires the video itself.

Online Appendix Page 5
We implement the supervised saliency map with Keras-vis\(^5\) and the unsupervised saliency map with the saliency module in OpenCV\(^6\). Figure B.1a presents an example of a video frame, and its corresponding frame in the supervised content engagement heatmap. Figure B.1b presents an example of an unsupervised content engagement map.

Figure B.1. An Example of the Content Engagement Heatmap

(a) Supervised Content Engagement Heatmap

(b) Unsupervised Content Engagement Heatmap

Note: The content engagement heatmap of a video is 3D. We present one frame of this 3D heatmap in this figure for illustration. A frame from the example video is shown in the left column. The corresponding frame in the content engagement heatmap is in the right column (supervised on the top, unsupervised at the bottom). Brighter areas in the content engagement heatmap correspond to pixels with higher saliency.

\(^5\)https://raghakot.github.io/keras-vis/vis.visualization.
\(^6\)https://docs.opencv.org/master/d8/d65/group.html.
C Details on the Product Placement Heatmap

We use SIFT to detect whether an advertised product appears in a given pixel of the video. We implement SIFT via ORB (Oriented FAST and Rotated BRIEF) in OpenCV.\(^7\)

Figure C.1 presents an example. The left column shows an image of the advertised product. The middle column shows a frame from the video. The dashed lines represent connections between the product image and the video frame that we detect using SIFT. These connections indicate the number and location of good keypoint matches. In most cases, the number will not be zero due to noise. A threshold is usually applied to filter out frames with false positive results. In the example frame, SIFT is able to correctly identify the product from the video despite substantial image rotation.

Following the ratio-test threshold of 0.75 explained in the paper, we assign binary values where 1 indicates that the product is detected at a given pixel and 0 indicates the opposite. The right column of Figure C.1 shows the corresponding frame from the 3D product placement heatmap of the video. The bright areas correspond to pixels where SIFT detects product presence.

Figure C.1. An Example of the Product Placement Heatmap

Note: We use SIFT to detect the product (left column) in a video frame (middle column). The corresponding frame in the 3D product placement heatmap of the video is shown in the right column, where the bright areas indicate product presence.

D Predicting Missing Product Category Information

One challenge we face in our cross-category analysis is that 68% of the products in our sales panel miss category labels. Our solution is to predict missing category labels based on product titles. To do so, we draw on a sample of 8,447 products with category labels (including products outside the sales panel to increase sample size). We assign 70% of products in this sample into the training set and perform cross-validation. We hold out the remaining 30% as the test set. We also make sure that the ratio of training to test data in each category is 70:30.

For pre-processing, we use packages quanteda,\(^8\) stopwords,\(^9\) and chinese.misc\(^10\) to tokenize the titles, delete stop words, and only keep the nouns. For feature extraction, we first construct a term-document matrix. Next, because titles from the same category often share common words, we use Latent Semantic Analysis (LSA),\(^11\) which measures word-word, word-passage, passage-passage relations by applying Singular Value Decomposition (SVD) to factorize the term-document matrix. Finally, we train the model with XGBoost\(^12\) in Caret.\(^13\) The model achieves 82% accuracy in the test sample. We have also tried ranger and rpart, achieving 63% and 79% accuracy, respectively. Based on predictive accuracy, we use the trained XGBoost model to impute missing category labels for products in our sales panel.

---

\(^8\)https://cran.r-project.org/web/packages/quanteda/quanteda.pdf.
\(^9\)https://cran.r-project.org/web/packages/stopwords/stopwords.pdf.
\(^10\)https://cran.r-project.org/web/packages/chinese.misc/chinese.misc.pdf.
\(^12\)https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf.
\(^13\)https://cran.r-project.org/web/packages/caret/caret.pdf.
E  Additional Summary Statistics of the Sales Panel

Figure E.1. Distribution of Video Engagement and Product Sales

Note: The sample consists of all videos/products in the sales panel, where each product corresponds to one video ad. The subfigures present, in order, the distribution of the number of likes, comments, and shares (video level), and average monthly sales revenue (product level), all in log scale.
Figure E.2. Distribution of the Number of Videos by Influencer

![Distribution of the Number of Videos by Influencer]

Note: The sample consists of all videos in the sales panel. Each observation is a video.

Table E.1. Summary Statistics of Influencer Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>St. Dev.</th>
<th>Min</th>
<th>Median</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>1,404</td>
<td>0.58</td>
<td>0.49</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td># Followers</td>
<td>1,404</td>
<td>1,617,806</td>
<td>3,048,990</td>
<td>0</td>
<td>723,679</td>
<td>43,012,100</td>
</tr>
<tr>
<td>Average Play</td>
<td>1,404</td>
<td>635,431</td>
<td>3,255,567</td>
<td>0</td>
<td>74,908</td>
<td>97,890,191</td>
</tr>
<tr>
<td>Expected CPM</td>
<td>1,404</td>
<td>1,026</td>
<td>21,314</td>
<td>0</td>
<td>121</td>
<td>785,714</td>
</tr>
<tr>
<td>Price per Video Ad</td>
<td>1,404</td>
<td>19,530</td>
<td>53,807</td>
<td>0</td>
<td>6,000</td>
<td>1,000,000</td>
</tr>
<tr>
<td># Video Ads Influencer Has Posted</td>
<td>1,404</td>
<td>13</td>
<td>26</td>
<td>0</td>
<td>2</td>
<td>265</td>
</tr>
</tbody>
</table>

Note: The sample consists of all influencers in the sales panel. Each variable is at the influencer level. All observations were recorded in January, 2019. For gender, 0 denotes female and 1 denotes male. CPM is the cost per 1,000 plays. Price per video ad is in RMB.
Figure E.3. Distribution of Video Length

Note: The sample consists of all videos in the sales panel. Each observation is a video.

Figure E.4. Distribution of Video Posting Date

Note: The sample consists of all videos in the sales panel. Each observation is a video.
Note: The sample consists of all products in the sales panel. Price is in RMB. Price range is between 5% and 95% quantiles. Dots represent mean prices in category.
Figure E.6. Engagement is Not Correlated with Sales

Note: The sample consists of the sales panel. Scatter plots of the relationship between engagement measures (the number of likes, comments, and shares) and product sales revenue. Both engagement and sales are residualized after controlling for product price, discount, video play volume, influencer fixed effects, and product category fixed effects. All variables are divided by 1,000 to facilitate presentation.
F User Survey of Product Category Perception

Figure F.1. Mean Ratings by Product Category
Callaway and Sant’Anna (2020) Specification

There are two steps in the estimation of an overall treatment effect as in Callaway and Sant’Anna (2020). First, a group-time average treatment effect is estimated for each group and time combination. Second, these group-time average treatment effects are aggregated over groups and time with proper weights.

A group-time average treatment effect (ATT) is defined as:

\[ \text{ATT}(g, t) = \mathbb{E}[Y_t(g) - Y_t(0) | G_g = 1], \]

where group \( g \) is defined as the set of units that are first treated at time \( g \), \( G_g = 1 \) indicates membership in group \( g \), and \( t \) denotes the calendar time.

To obtain an overall treatment effect, we first aggregate over periods, where \( T \) is the terminal period:

\[ \theta(g) = \frac{1}{T - g + 1} \sum_{t=g}^{T} \text{ATT}(g, t), \]

then aggregate over groups:

\[ \theta = \sum_{g=1}^{G} \mathbb{P}(G = g | G \leq T) \theta(g), \]

where \( \mathbb{P}(G = g) \) is the group size (the fraction of units in a given group).

We estimate the overall treatment effect \( \theta \) separately on the full sample and a subsample with videos in the top 10% of the m-score distribution. The computation is performed using the DID package in R.\(^{14}\)

\(^{14}\)https://cran.r-project.org/web/packages/did/index.html.