Engagement that Sells: Influencer Video Advertising on TikTok[†]

Jeremy Yang, Juanjuan Zhang, Yuhan Zhang

September 6, 2022

Abstract

Many ads are engaging yet ineffective. This problem is exacerbated in influencer advertising when the incentives of influencers and advertisers are not perfectly aligned. This paper develops an algorithm to predict the effect of influencer video advertising on product sales. We propose the concept of product engagement score, or pe-score, to capture how engaging the product is as presented in an influencer video ad. We locate product placement with an object detection algorithm, and estimate pixel-level engagement as a saliency map by training a deep 3D convolutional neural network on video-level engagement data. Pe-score is computed as the pixel-level, engagement-weighted product placement in a video. We construct and evaluate the algorithm with around 20,000 influencer video ads on TikTok. We leverage variation in video posting time to identify the causal effect of video ads on product sales. Videos with higher pe-scores indeed lift more sales. This effect is robust and more pronounced among impulsive, hedonic, or inexpensive products. Meanwhile, engagement increases with human presence, sad or happy emotions, and stimulating or novel activities. We discuss how various stakeholders in influencer advertising can use pe-score in a scalable way to develop content, align incentives, and improve efficiency.

Keywords: influencer advertising, video advertising, entertainment commerce, content strategy, sales conversion, incentive alignment, computer vision, TikTok.

[†]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. The authors thank each other for sharing many hardships of the COVID-19 pandemic and a challenging job market. The authors thank Kun Cheng and Haiwen Li for research assistance, and Yuchen Wang for participation in early-stage data analysis. The authors received helpful comments from faculty and students in the MIT Sloan Marketing Group; from seminar participants at Cheung Kong Graduate School of Business, City University of Hong Kong, Columbia University, Erasmus University, Harvard University, HEC Paris, Massachusetts Institute of Technology, Meta, Northwestern University (QME Rossi Seminar), Peking University, Purdue University, Temple University, TikTok, University of Hong Kong, University of Houston, University of Toronto, University of Virginia, and University of Washington; and from attendees of the 2020 Conference on Digital Experimentation, 2020 HBS Digital Doctoral Workshop, 2020 NYU-Temple-CMU Conference on Artificial Intelligence, Machine Learning, and Business Analytics, 2021 Annual Conference of JMS China Marketing Science, 2021 Artificial Intelligence in Management Conference, 2021 ISPOC Seminar, 2021 Marketing Science Conference, 2021 MIT IDE Annual Conference, 2021 Paris Conference on Digital Economics Workshop, 2021 Theory + Practice in Marketing Conference, 2022 HBS Data Science in a Digital World Conference, and 2022 MIT CMO Summit.

1 Introduction

"The next Amazon competitor is going to look like a social or video app, not a shopping app," says Connie Chan, a 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 platform, 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" or "social 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 in the world since 2018, TikTok has over 1.8 billion active users around the globe and is projected to reach 3.3 billion downloads by the end of 2022.[3] E-commerce integration is prominent on TikTok, especially in its original Chinese version. An ecosystem has been developed where product sellers routinely pay influencers to place products in their videos, while users make purchases using product links in the videos. At its rate of growth, TikTok is expected to bring in more influencer marketing spending than Facebook in 2022 and YouTube by 2024.[4]

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.[5] As a result, sellers

[1] 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.

[3] TikTok revenue and usage statistics. *Business of Apps*, June 30, 2022.

[4] TikTok to overtake Facebook in influencer marketing spend this year, YouTube by 2024. *TechCrunch*, August 2, 2022.

[5] The company that provided us data emphasizes strong industry demand for such predictive tools.

often rely on influencer engagement metrics (such as the number of likes, shares, and comments) for campaign management. On TikTok, many sellers would simply choose an engaging influencer, then 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.[6]

The goal of this paper is to develop a method to predict the effect of influencer video ads on product sales. Our argument is that ads can be engaging for the "wrong" reason – what makes them engaging may have little to do with the advertised product. Influencer advertising can be particularly susceptible to this problem, because influencers are often incentivized to promote their personal brand, not just the product.[7] As such, they may not want to allocate the most engaging space and time of their videos to the product, which lowers ad effectiveness. Based on this argument, we develop a metric that captures to what extent engagement is driven by the product, i.e., being engaging for the right reason. We call this metric *product engagement score*, or *pe-score* for brevity.

We operationalize pe-score so that it is intuitive, is able to turn unstructured video data into structured information, and is measurable prior to ad release for better campaign management. To meet these objectives, we define a video ad's pe-score as the average pixel-level engagement score over the pixels in which a product is presented. We compute pe-score in three steps.

First, we construct a three-dimensional (3D) *engagement heatmap* for each video to measure the contribution of each pixel to overall video engagement. The three dimensions are the height and width of each video frame in pixels and the length of the video in seconds. We train a deep 3D convolutional neural network using video-level engagement data. A video's engagement heatmap is then derived as a pixel-level saliency map, which outputs the gradient of video-level engagement with respect to each pixel in the video.

[6] One million likes but less than 5,000 monthly sales, what did the product do wrong in short video marketing? *CAAS Data*, May 17, 2020.

[7] B2B influencer marketing research shows disconnect between brands and influencers. *OST*, January 2, 2019.

Second, we construct a 3D *product heatmap* for each video that has the same dimension as the engagement heatmap. The product heatmap shows whether the advertised product is present at a given pixel in a given frame of the video. We estimate the product 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."

Third, we compute pe-score as the Frobenius inner product of the two 3D matrices, normalized by the total number of pixels of the video. Pe-score can thus be interpreted as a video's engagement level over the pixels in which a product is presented, or in other words, how engaging the product is as presented in a video ad.

We hypothesize that a video ad with a higher pe-score is more effective in lifting sales. Note that a video ad that is engaging overall or features the product throughout does not necessarily have a high pe-score. In the former case, overall engagement might be driven by non-product content; in the latter case, product presentation might be uninteresting. A high pe-score, as its name emphasizes, requires the product itself to be engagingly presented in a video ad.

We evaluate our method using a proprietary dataset of around 20,000 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 May to November 2019.[8] 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. For a smell test of our incentive-misalignment argument, we also collect an auxiliary dataset, in which influencers advertise their own products. Consistent with our argument, pe-score tends to be higher in these videos than in videos where influencers advertise for another party.

For our main test, we estimate the causal effect of influencer video ads on product sales via the staggered difference-in-differences (diff-in-diff) method, leveraging the vari-

[8]Owned by Alibaba, Taobao is one of the world's largest e-commerce websites and the major platform on which products advertised in TikTok influencer videos were sold during the time of our data.

ation in video posting time for identification. Consistent with our hypothesis, pe-score positively moderates the sales lift of a video ad. A 1% increase in pe-score is associated with a 2.58% increase in sales revenue of the advertised product. Notably, overall video engagement alone has no moderation effect on sales lift. Meanwhile, product placement has a negative moderation effect, suggesting that unilaterally increasing product appearance without improving engagement might actually hurt sales compared with not advertising at all. Morever, being both engaging and intensive with product placement but doing so separately has no moderation effect either. These results highlight the unique predictive power of pe-score – simply making the video more engaging or simply featuring the product more does not help; it is product engagement that drives sales.

Our findings are robust with respect to alternative ways to construct pe-score and to identify its sales impact. To better understand the applicability of pe-score, we conduct a supplementary survey to classify advertised products in our data. Pe-score is more predictive of sales in low-involvement 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 possible drivers of pe-score. Leveraging the engagement heatmap and a series of computer vision algorithms, we find that pixel-level engagement increases with human presence, sad or happy facial expressions, and stimulating or novel activities. It may be helpful to align product placement spatiotemporally with these elements of engagement.

Pe-score can be practically valuable in several ways. First, we invested heavily in algorithm calibration, such that pe-score can be easily computed for a video ad in future applications, the only data requirement being the video ad itself.[9] This means influencers can use the algorithm as an automated tool to test their videos in the creative process prior

[9]We also need an image of the advertised product to construct the product heatmap. It can be a product image from e-commerce websites but can simply be a cropped image from the video when the product is shown.

to release. Second, pe-score introduces a new contractual instrument to the influencer advertising space. Sellers can use pe-score to screen candidate videos or directly write a contract based on it. Influencers can use pe-score to signal their business involvement beyond what engagement metrics are able to communicate. Platforms can design various policies to use pe-score for more accurate attribution and more efficient allocation. After all, the pe-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 $16 billion industry in 2022 with a whopping 30% growth rate.[10] 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. With its rapid growth, influencer video advertising is gaining attention

[10] The state of influencer marketing 2022: benchmark report. *Influencer Marketing Hub*, March 2, 2022.

in academia. One notable recent study is Rajaram and Manchanda (2020), which 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 ad engagement versus conversion. Teixeira et al. (2014) found that ad entertainment increases viewing but has an inverted U-shaped effect on purchase intent. Tucker (2015) showed a negative relationship between ad views and purchase intent, an effect likely driven by outrageous video ads purposely designed to provoke sharing. John et al. (2017) found that liking a brand on Facebook has no positive impact on consumer attitudes or purchases; what matters is consumers' pre-existing preference for the brand. Using actual sales data, we also find that engagement does not guarantee conversion. Furthermore, we propose and validate a novel metric that connects engagement with conversion. We find that what drives conversion is not engagement per se, but effective engagement related to the advertised product.

Our paper is also related to the marketing literature on video content design.[11] One prominent stream of research links video content with moment-to-moment (MTM) viewer behaviors, meaning behaviors that vary over time during the process of consuming a video. 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,

[11] Another related, burgeoning line of research uses image data to inform various aspects of marketing, including social media engagement (Li and Xie 2020), brand image extraction (Liu et al. 2020), facial image mining (Tkachenko and Jedidi 2020), product aesthetics (Burnap et al. 2021), brand selfies (Hartmann et al. 2021), listing-image design (Zhang et al. 2021), logo design (Dew et al. 2022), and business survival prediction (Zhang and Luo 2022).

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 ads. Indeed, we find that influencers' incentive to promote themselves does affect ad 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 instead 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 prior to release.[12] 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 has 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 pe-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.[13] However, purchase interest can be activated in the process of consuming a video ad, if the advertised product happens to grab user attention. Based on this idea, our hypothesis behind pe-score is that, other things being equal, the more engaging an advertised product is in an influencer video ad, the more

[12] Our attention to scalable video analysis echoes Li et al. (2019), one of the first video-mining papers in marketing. Their paper advocated the use of visual variation and video content, measures that can be automatically extracted from videos to explain crowdfunding outcomes.

[13] TikTok ads: Everything you need to know about marketing on TikTok. *Oberlo*, November 21, 2020.

effective the video ad will be in lifting sales.[14] To operationalize this idea, we propose a three-step algorithm:

1. Compute a pixel-level engagement heatmap over the video ad to identify the most engaging spots of the video.

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

3. Compute pe-score as the normalized inner product of the two heatmaps to capture the average product engagement of the video.

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

3.1 Engagement Heatmap

For each video ad, we first estimate an engagement heatmap, which is a 3D array that captures the spatiotemporal variation of content engagement in the video. The three dimensions of the 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 recognition (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, Xception, Chollet 2017) with the top classification layer removed,[15] then feeding the feature sequence into a 3D convolution layer which accounts

[14] The seminal paper of Mitchell and Olson (1981) found that consumers' attitude towards an ad can mediate their attitudes 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 engaging way.

[15] ImageNet (http://www.image-net.org) is a database of over 1 million images with 1,000 class labels. It is considered the industry standard for training and testing image classification algorithms. Xception is an effective network for image classification, with a top-1 accuracy of 0.79 and top-5 accuracy of 0.95.

for the temporal dependencies across frames (e.g., Tran et al. 2015).

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 context 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.

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 on their social networks.[16] However, our results are robust if we use likes to measure engagement (see Online Appendix I.1).

Prior to training, we regress video-level raw engagement data on influencer fixed effects, product fixed effects, acoustic features, and transcript embeddings. We retain the residuals from the regression and use them as the output to train the 3D CNN. Using engagement residuals instead of raw engagement allows us to focus on the remaining variation in engagement that is driven by the visual component of the video ad holding other features that might affect engagement fixed.[17] In the rest of the paper, video-level engagement refers to this "residualized" engagement value unless otherwise noted.

We focus on videos with spoken words so that a valid transcript can be extracted, although our results are robust if we relax this requirement (these results are in an earlier version of the paper). As we will detail in Section 4, the sample we rely on to construct our algorithm contains 16,951 video ads. We train the 3D CNN on 10,000 videos, validate it on 3,500 videos, and test it on 3,451 holdout videos.

To appreciate the magnitude of the raw data for pixel-level analysis, consider a typ-

[16] Social media metrics compared: Which are the most valuable? *Social Media Week*, October 19, 2017.

[17] For example, it is important to note that pe-score measures the engagement of product placement or presentation rather than how engaging the product itself is by design. Product design is fixed by the time influencers are making video ads. Product fixed effects help control for the possibility that some products are more engaging by design.

ical TikTok video. It is most commonly 15-60 seconds in length and has up to 60 frames per second (fps). 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* 15-second, 60fps TikTok video would contain 15 * 60 * 1,080 * 1,920 * 3 = 5,598,720,000 pixel values. To make the training process tractable, we sub-sample videos to one frame per second and resize each frame to a dimension of 224p * 224p.[18] This allows each video to be represented as a much more feasible (S, 224, 224, 3) numerical array, where S is the duration of a video in seconds.

In the end, our full 3D CNN has over 7 million trainable parameters, over 2 million input variables (with each pixel value at a given color channel being an independent variable), and takes in over 20 billion data points in the training process.[19] We train the 3D CNN with graphics processing units (GPUs) on a high-performance computing (HPC) cluster using TensorFlow.[20] It achieves an accuracy of 73% (one minus the mean absolute percentage error, or MAPE) on the test set. See Online Appendix A for more details on the network structure and the training process.

After training the 3D 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 while holding other pixel values fixed. For images with color, we follow the common practice to compute three partial derivatives for each of its RGB color channels at a given pixel and take the maximum of the three as the saliency value at that pixel. The magnitude of the derivatives tells us how much the outcome variable, video-level engagement,

[18]This is the standard image size for many widely used computer vision algorithms. However, our algorithm should accommodate any image size in principle.

[19]There is a paper famously titled "I Just Ran Four Million Regressions" (Sala-i Martin 1997). Here we just ran one regression with over four million parameters.

[20]https://keras.io.

changes with respect to changes in pixels of the input image, or equivalently, which pixels need to be changed to affect video-level engagement the most. We interpret a high absolute value of the derivative as high engagement at that particular pixel.[21] These inferred pixel-level engagement values, or attention, in saliency maps have been shown to accurately predict the actual gaze map based on eye tracking data (Borji et al. 2013, Dupont et al. 2016).

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 entire video. This allows us to capture any dependency across video frames when deciding which pixels are driving engagement. We estimate saliency maps using the trained 3D CNN with tf-keras-vis.[22] See Online Appendix B for more details on the saliency map.

To summarize the engagement heatmap from the econometric perspective, we use one video-level engagement measure (the number of shares) to back out engagement scores (the partial derivatives) at each pixel in the video. This is analogous to a regression of video-level engagement on all pixel values in the video, except that our algorithm can handle high-dimensional inputs and arbitrary correlation between pixel values. Similar to standard regressions, this approach requires sufficient spatiotemporal variation in pixel values. This is likely satisfied given the high variety of videos on TikTok. To be able to interpret pixel-level engagement as causal effects, we also need the "no design endogeneity" assumption, meaning there are no omitted factors driving both pixel values and video-level engagement. It would be concerning if, for example, a bright video is engaging because a cheerful influencer uses bright pixels and engages the audience with a bright personality. Our use of residualized video-level engagement mitigates this concern by controlling for influencer and product fixed effects as well as acoustics and

[21] We ignore the sign because increasing or decreasing the value of a pixel along a particular color channel simply means changing its intensity or brightness (see http://www.cknuckles.com/rgbsliders.html for an interactive example), which does not have an intrinsic, directional meaning.

[22] https://github.com/keisen/tf-keras-vis.

spoken content. Lastly, like standard regressions, estimation relies on functional-form assumptions. However, the functional form is not imposed a priori but learned from the data via a flexible 3D CNN.

3.2 Product Heatmap

For each video, we estimate a product heatmap, which is a 3D array of the same dimension as the 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.[23]

SIFT is a popular algorithm for object detection, matching features across different images to identify the presence of an object in 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 or query 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.[24]

Because SIFT matches at the pixel level, the identified product pixels can be scattered in a video frame and do not necessarily enclose the entire product. We connect these pixels to create a convex hull and consider all pixels within the convex hull as product pixels. The resulting product heatmap is a 3D array of binary values, where 1 indicates

[23] https://docs.opencv.org/master/dc/dc3/tutorial/pymatcher.html.

[24] 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, the best match is more likely to be noise. If the two are different enough, 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.

product presence at a pixel and 0 indicates absence. See Online Appendix C for more details on the product heatmap.

3.3 Computing PE-Score

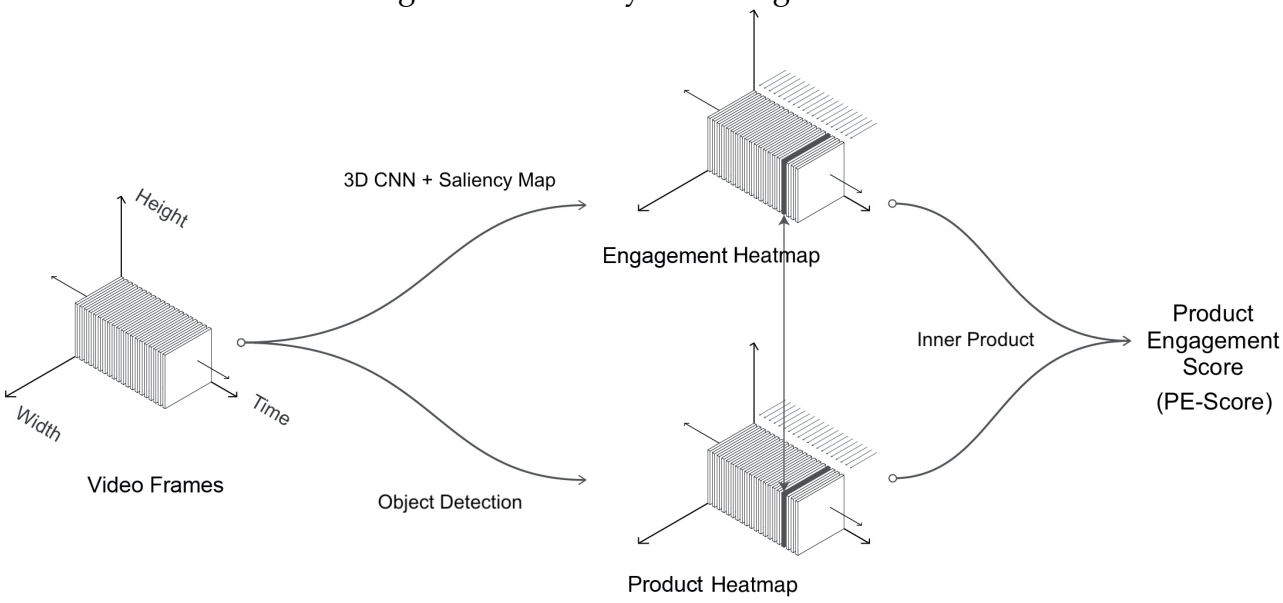
In the third step, we combine the engagement heatmap and the product heatmap to calculate pe-score. Let e_{hwsv} be the (continuous) pixel-level value in the 3D engagement heatmap and p_{hwsv} be the (binary) pixel-level value in the 3D product heatmap. The symbols h, w, s, and v index height (in pixels), width (in pixels), time (in seconds), and video, respectively. We define a video's pe-score as the normalized inner product of the two heatmaps:

$$pe_v := \frac{1}{H_v W_v S_v} \sum_{h,w,s} e_{hwsv} \cdot p_{hwsv}, \tag{1}$$

where H_v, W_v, and S_v are the total height, width, and length of video v, respectively. As explained, we reshape the frame size for each video so that H_v and W_v are the same across videos, while we allow S_v to be different across videos. Their product, $H_v W_v S_v$, is thus the total number of pixels, or the volume, of video v. As discussed, we interpret pe-score as the average product engagement of a video; the inner product captures the spatiotemporal synchronicity between content engagement and product placement. We summarize the algorithm in Figure 1.

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) train the algorithm using a large sample of videos, (2) use transfer learning to reduce the number of parameters to estimate, (3) use effective regularization methods such as dropout (Srivastava et al. 2014) and early-stopping, and (4) check for overfitting on the validation sample. Reassuringly,

Figure 1. Summary of the Algorithm



as we discuss in Online Appendix A, test result suggests 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 lift as the outcome variable. There are three reasons. First, CNN results are typically difficult to interpret (Zhang and Zhu 2018). We compute pe-score as an interim summary statistic that is succinct, behaviorally meaningful, and interpretable. Theoretically, this allows us to achieve greater conceptual clarity on *what* lifts sales. Practically, knowing what lifts sales sheds light on the famously challenging problem of marketing attribution (e.g., Testwuide 2020). Second, we do not observe the sales *lift* of each video ad in the data; it needs to be estimated via a diff-in-diff approach, as we will explain. These estimates at the video level can be noisy. Instead, we will estimate a single average sales lift and examine the moderation effect of pe-score. Third, the number of videos with sales data (2,685; see next section) may not be sufficient to train a complex 3D CNN. We leave this analysis as a worthy topic for future research.

In what follows, we test our algorithm with sales data. Specifically, we test whether influencer video ads of high pe-score lift more sales. We present the data next.

4 Data

We test whether our algorithm can help predict sales lift. To do so, we need data on influencer video ads, video engagement metrics, and sales of the advertised products. We are fortunate to have developed such a dataset via collaboration with an entertainment commerce company. For context, at the time of this study, content and engagement data were usually stored in one system (i.e., social media platforms such as TikTok), while sales data were typically stored in another (e.g., e-commerce websites such as Taobao). It is valuable to be able to connect these data sources.[25]

We collect influencer video ads data from the Chinese version of TikTok because of its mature ecosystem around influencer video advertising. There is an established marketplace called Xingtu, where sellers contract with influencers to advertise their products. To date, Xingtu has attracted about 1.5 million influencers and 1.2 million registered sellers.[26] Two notable features characterize this marketplace. First, engagement is the centerpiece of the ecosystem. It determines how influencers price their video ads, how sellers search for influencers, and how sellers monitor ad performance. Second, influencers have significant discretion in designing their video ad content. In a typical ad creation process, a seller provides some general guidelines, 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. Sellers are able to influence ad content to some degree. However, there are many video design aspects that are controlled by the influencer. In particular, there is no clear way for sellers to predict sales lift from an ad. They pay for engagement, in the (sometimes shattered) hopes that engaging influencers would lift sales.

To further understand the TikTok influencer advertising market, we interviewed a number of practitioners in this space. Online Appendix D presents the scripts. These interviews suggest that, indeed, (1) sellers do not tend to influence the visual aspect of

[25] For instance, Lee et al. (2018) studied advertising content and engagement on Facebook, where the lack of access to sales data was a limitation.

[26] Xingtu's official website.

video content that we focus on in the paper, (2) sellers do not tend to influence product placement, and (3) influencers do not tend to choose the posting time of video ads based on product-specific demand.

We capture sales data on Taobao. Taobao is the biggest e-commerce website in China.[27] The vast majority of sellers in our video ads data list their products on Taobao exclusively, as indicated by the product link in the video ads. We also confirmed with our partner company that TikTok and Taobao were indeed the main advertising and sales channels for sellers during the time of our sample. This helps us attribute product sales lift on Taobao to video ads on TikTok.

More specifically, our dataset is a matched sample from two separate sources. The first is a video dataset that contains all TikTok influencer video ads with product links from March to June 2019. For each influencer ad, the data contain the video, its posting date, product ID, corresponding engagement metrics measured at the daily level, as well as influencer characteristics.

The second source is a product dataset that contains all products listed on Taobao from May to November 2019. For each of these products, the data track its product ID, sales revenue on Taobao, product image, category, price, and discount. Product revenue is summed over the previous 30 days, including the current day, measured at the daily level.[28] Some products are missing category information. We use product titles and 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 E for details). The majority of products in the data have one video ad. We focus on these products in subsequent analysis for clean attribution.[29]

[27] Top 15 Chinese E-commerce websites in 2022. *TMO Group*, April 6, 2022.

[28] There are some missing sales observations for technical reasons that are believed to have occurred randomly.

[29] 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."

We match the two data sources using product ID.[30] Among influencer video ads from the first source, 2,734 have matching product sales data. Among the 2,734 video ads, 2,685 have complete influencer characteristics. We call the sales panel dataset of these 2,685 product-video pairs the *sales panel*, which we will hold out and use to test our algorithm's ability to predict sales lift. Among the remaining videos from the first source, as discussed, we focus on videos with spoken words to control for video transcript. This yields 16,951 video ads that we use to construct the algorithm: 10,000 for training the 3D CNN, 3,500 for validation, and the remaining 3,451 for holdout test, all through random assignment. Once the algorithm is constructed, pe-scores of videos in the sales panel can be computed without relying on their engagement or sales data (i.e., prior to release). This feature contributes to algorithm scalability. Meanwhile, it leaves us with a high-power test of our algorithm – we construct it from one sample of videos, and test its predictive power on sales lift on a different sample, which helps examine the external validity of the algorithm.[31] Altogether, this paper draws on a total of 19,636 video ads.

Table 1 presents summary statistics of video engagement metrics: the number of likes, comments, and shares.[32] We do observe video engagement at the daily level. However, engagement takes time to grow. To capture each video's ultimate level of engagement, 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 were shared less on average.

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

[30] Some products might have changed their ID during the data collection, which prevented us from matching every video to the corresponding product.

[31] The sales panel may not be a random subset of products on Taobao. We expect the effect magnitude of pe-score to depend on the specific product market. However, we constructed our algorithm based on all available TikTok influencer video ads that have spoken words during our data window. Therefore, our *algorithm* can plausibly generalize.

[32] 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.

Table 1. Summary Statistics of Video Engagement Data

Variable	N	Mean	St. Dev.	Min	Median	Max
Videos in the Sales Panel						
Likes	2,685	38,515	111,116	0	3,654	1,831,709
Comments	2,685	542	2,052	0	84	71,068
Shares	2,685	936	5,007	0	80	166,821
Videos for Algorithm Construction (Training, Validation, and Test Sets)						
Likes	16,951	34,339	112,302	0	3,021	2,553,627
Comments	16,951	531	2,124	0	69	71,068
shares	16,951	1,184	6,690	0	91	195,563

Note: Each engagement metric is at the video level.

246,680 RMB, or 35,699 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)

Variable	N	Mean	St. Dev.	Min	Median	Max
Average Sales Revenue	2,685	246,680	5,288,389	0	9,446	272,107,695
Price	2,685	1,081	39,220	0	68	2,019,515
Discount	2,685	100	506	0	20	13,901

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 F. In summary, engagement and sales show sizable variation across videos (Figure F.1). Most influencers post one video ad, although there is a distribution (Figure F.2). This variation allows us to control for influencer fixed effects in subsequent analysis. Nevertheless, we report influencer summary statistics in Table F.1. In addition, the most common video length is 15 seconds whereas video posting date is widely distributed, a fact we will leverage for causal identification of sales lift (Figure F.3). 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 F.4).

A particularly revealing pattern in the data is the lack of significant correlation be-

tween video engagement and product sales. Figure F.5 presents the scatter plots of the relationship between raw engagement metrics (the number of likes, comments, and shares) and product sales averaged over their observed days in our sample. Sales has no significant correlation with the raw engagement metrics (ρ = 0.0022, p = 0.91 for likes; ρ = 0.0024, p = 0.90 for comments; ρ = 0.00093, p = 0.96 for shares). This result suggests that using video engagement to evaluate ad effectiveness can be misleading. Our algorithm is intended to address this problem. We present its evaluation in the next section.

5 Algorithm Evaluation – Main Results

In this section, we first present the computational results of our algorithm. We also show suggestive evidence of the incentive misalignment argument. We then proceed to the main test, of whether influencer video ads with higher pe-scores lift more sales. Lastly, we explore whether the effect is stronger in some categories than others.

5.1 Computational Results of the Algorithm

For each video ad, the algorithm outputs a 3D engagement heatmap, a 3D product heatmap, and a pe-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 score, termed to differentiate it from actual engagement, is a video's sum of pixel-level engagement values, and computed product score 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.

To further visualize these computational results, we present average pixel-level engagement and product scores 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

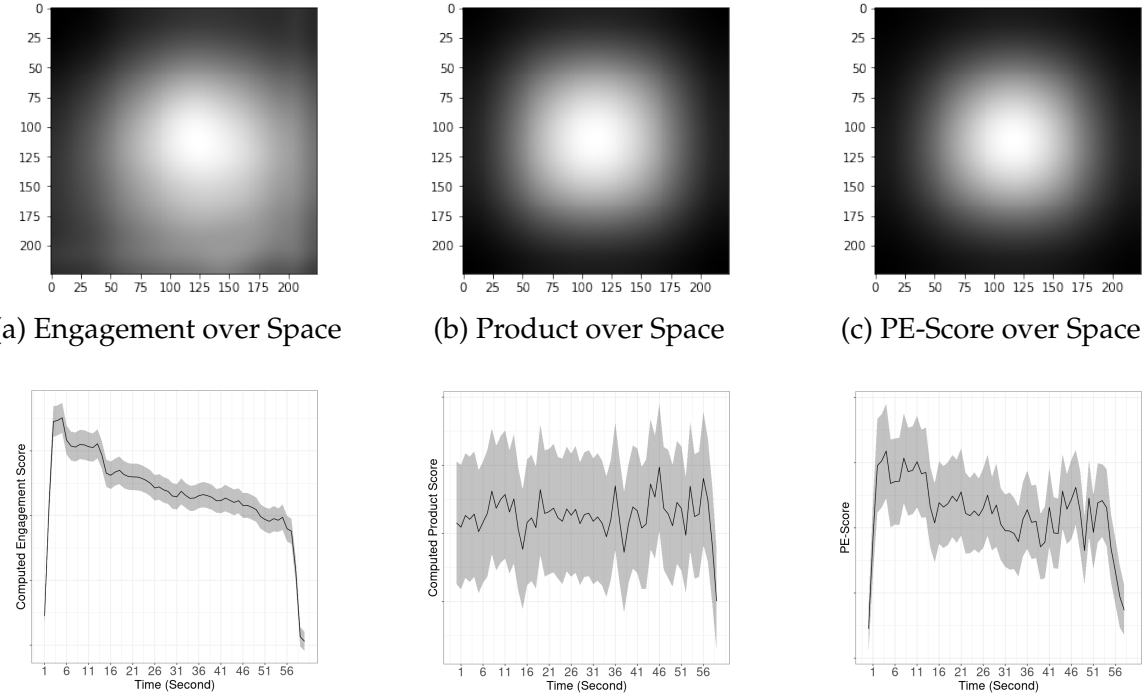
Table 3. Summary Statistics of Video-Level Computed Engagement Score, Product Score, and PE-Score

Variable	N	Mean	St. Dev.	Min	Median	Max
Computed Engagement Score	2,685	0.48	0.15	0.00	0.49	1.00
Computed Product Score	2,685	0.18	0.14	0.00	0.15	1.00
PE-Score	2,685	0.21	0.14	0.00	0.19	1.00

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 scores and product scores are aggregated from their pixel-level values.

pe-score, computed as the pixel-level product of engagement and product score, averaged over either space (Figure 2c) or time (Figure 2f).

Figure 2. Distribution of Computed Engagement Score, Product Score, and PE-Score



(a) Engagement over Space (b) Product over Space (c) PE-Score over Space

(d) Engagement over Time (e) Product over Time (f) PE-Score over Time

Note: For subfigures (a)-(c), brighter means higher engagement, more product placement, and higher pe-scores, respectively. For subfigures (d)-(f), gray areas represent values within 0.1 standard deviation from the mean.

We see a pattern. The most engaging region of a video frame tends to be its center, with the bottom and right side of the screen being slightly more engaging than the top and left. This is possibly because the bottom is where the information of the video (e.g.,

influencer name, a short description and hashtags of the video) and the right is where engagement metrics are shown. Similarly, on average, products tend to appear in the center of the frame and pixel-level pe-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 of a video vary from frame to frame. The average structural similarity index measure (SSIM) between two consecutive engagement heatmaps in our data is 0.78.[33] In fact, products do not always appear in high engagement pixels. The SSIM between the engagement heatmap and the product heatmap on the same frame averaged over all videos in the sales panel is 0.46, which is moderate.

Over the duration of a video, engagement tends to start low in the first second, rise rapidly and peak in the first 6 seconds, decline gradually from the 7^{th} to 57^{th} seconds and very sharply in the last 3 seconds. Product placement is noticeably different; it tends to be uniform except in the last 3 seconds. Pe-score follows a pattern similar to engagement – it rises then falls and falls sharply 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 areas in the figures represent values within 0.1 standard deviation from the mean, which span a noticeable range already. These observations again highlight the incremental value of our algorithm, which captures rich heterogeneity across space, time, and videos. In the next section, we use the computed pe-scores to show evidence of incentive misalignment, the argument underlying our algorithm.

5.2 Evidence of Incentive Misalignment

Our algorithm is based on the argument of influencer incentive misalignment. At the time of this study, influencers are typically paid a fixed price per video ad which is mostly

[33] SSIM is a value between 0 and 1 that measures the perceived similarity between two images. It takes additional contextual information such as luminance and contrast into account compared to measures such as Pearson correlation or mean squared error (MSE).

driven by the number of followers and engagement. Influencers may thus have more incentives to optimize a video ad for engagement rather than sales lift. Meanwhile, product ads during entertainment are generally disliked (e.g., Elpers et al. 2003, Wilbur 2016); influencers may even lose followers by posting sponsored videos (Cheng and Zhang 2022). In light of the pe-score concept, this means influencers may avoid placing the product in the most engaging spots of the video. It is important to note that, even though sellers can fully observe a video ad, they may have different *interpretation* of the content than the influencer. For example, sellers may not know what engages a particular influencer's followers. This is similar to how medical notes or legal documents might be fully observable to both doctors and patients or lawyers and clients, but information asymmetry could still arise due to the asymmetry in knowledge. This information asymmetry is also a reason why sellers ask influencers to design the video ads in the first place.

We supplement this discussion with a smell test of the incentive misalignment argument. We collected a separate sample of 77 video ads, where influencers advertise their own products, and compare them with the 2,685 videos in the sales panel. If the incentive argument is true, these 77 video ads should have higher pe-scores than those in the sales panel.[34] We pool these two types of videos and regress pe-score on an indicator variable of whether the influencer is advertising her/his own product, while controlling for product price, discount, and influencer characteristics. We find that pe-scores are, on average, 28% higher when influencers are advertising for their own products.

Our argument is that influencers are *able* to design effective video ads but may act differently for strategic reasons. To further support this argument, we examine the effect of influencer experience. If the lack of ability, as opposed to incentive, is what hinders advertising effectiveness, influencers with less experience should produce lower pe-scores. We regress pe-score on measures of influencer experience, including the number of video

[34] 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 (2022) for a theory in which influencers receive financial incentives to review products.

ads the influencer has posted and the number of days since the influencer's first post, controlling for other influencer characteristics (see Table F.1).[35] We find no statistically significant association between pe-score and these influencer experience measures, contrary to the ability explanation.

Taken together, the evidence is consistent with the incentive misalignment argument. Our algorithm can help mitigate this problem by quantifying to what extent the influencer is effectively advertising the product. We test our algorithm in the following section.

5.3 Influencer Video Ads with Higher PE-Scores Lift More Sales

We begin by presenting the model-free relationship between pe-score and difference in sales revenue before and after a video ad is posted. We calculate each product's sales difference as its average sales revenue within the data window after the posting of its video ad minus its average sales revenue before. We plot the sales difference against the pe-score of the corresponding video for products in the sales panel. Figure 3 shows the scatter plot. There is a positive correlation between pe-score and sales difference ($\rho = 0.16, p = 0.01$), consistent with our main hypothesis.

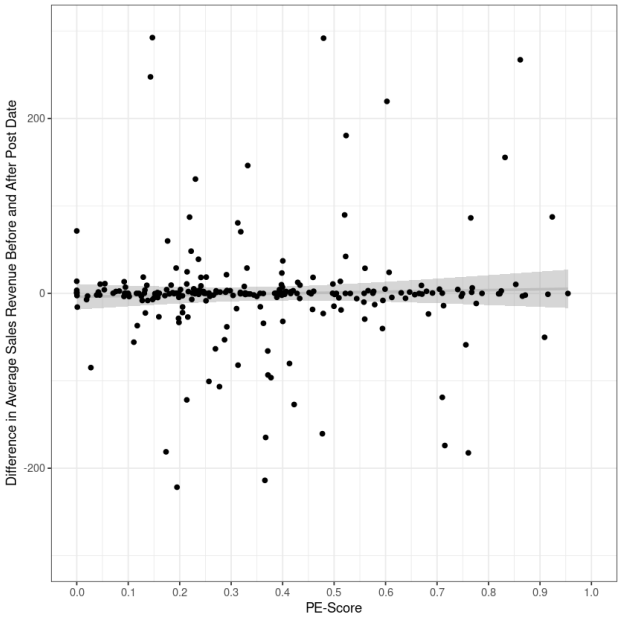
We formally test our main hypothesis in two steps. We first quantify the causal impact of influencer video ads on sales. We then examine whether this causal sales lift is stronger among influencer video ads with higher pe-scores. We identify the causal impact of influencer video ads using the staggered diff-in-diff approach, estimating the following specification:[36]

$$Sales_{vd} = \alpha \cdot Post_{vd} + Video_v + Influencer_{i(v)} + Day_d + Search_{vd} + \epsilon_{vd}. \quad (2)$$

[35] 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 F.1, we do not include this characteristic as a moderator variable of advertising effect.

[36] We present the standard two-way fixed effects regression in the main text. Given the recent literature on the potential problems with this approach when treatment timing is staggered and treatment effects might be heterogeneous across units or over time, we run diagnosis tests and alternative specifications, including two-stage diff-in-diff (Gardner 2022) and imputation diff-in-diff (Borusyak et al. 2022), as robustness checks. See Online Appendix I.7 for details.

Figure 3. Before-After Sales Difference by PE-Score



Note: Sales difference on y-axis equals a product's average sales revenue (in 1,000 RMB) after the posting of its video ad minus its average sales revenue before. Each dot is a video/product. We restrict y-axis to values between -300 and 300 for visualization purpose and it retains over 90% of the sample. The regression line shows an upward trend ($\rho = 0.16, p = 0.01$). Gray areas represent the 95% confidence band along the regression line.

The dependent variable, $Sales_{vd}$, is the prior-30-day sales revenue of a product v as measured on day d.[37] As discussed, we focus on products that have only one video ad, so that v indexes both the video and the product. Meanwhile, recall that influencers vary in the number of videos they post, and so we use the mapping $i(v)$ to index influencers separately from videos.

The dependent variable is cumulative sales, but we rely on staggered diff-in-diff to identify incremental sales at the daily level. To fix ideas, consider a treated product v and a control product u. Let $d = 1$ be the day of treatment and $d = 0$ the day before treatment. The diff-in-diff estimator is $(Sales_{v1} - Sales_{v0}) - (Sales_{u1} - Sales_{u0})$. By definition, $Sales_{vd}$ is product v's cumulative sales from day $d - 29$ through day d. Therefore, $Sales_{v1} - Sales_{v0}$ is sales on day 1 minus sales on day -29 for the treated product. Similarly, $Sales_{u1} - Sales_{u0}$ is sales on day 1 minus sales on day -29 for the control product, which measures the time

[37] Our results are robust if we use log-transformation of sales as the dependent variable.

effect on sales between day -29 and day 1 in the absence of the treatment. Therefore, under the standard parallel-trend assumption, which we show to hold in Online Appendix I.3, the diff-in-diff estimator identifies the treatment effect on incremental sales on day 1, the day of treatment.

The treatment variable is $Post_{vd}$, which equals 1 if video ad v is posted by day d and equals 0 otherwise. The parameter α measures the average sales revenue lifted by an influencer video ad. Because the dependent variable is prior-30-day sales, we examine an alternative specification in which the treatment variable is prorated based on the number of days an ad is present out of the 30-day window; results are largely robust (Online Appendix I.4). We also run a simulation study and find that $Post_{vd}$ captures the true treatment effect reasonably well (Online Appendix I.5).

Leveraging the panel structure of the data, we include video/product fixed effects $Video_v$ and influencer fixed effects $Influencer_{i(v)}$ to control for unobserved heterogeneity across videos/products and influencers, respectively. We also include day fixed effects Day_d to capture common time effects (e.g., trends, seasonality) on sales. As mentioned, the rich variation in video posting date allows us to separately identify the treatment effect of video ads from these common time effects.

Reverse causality could be a concern if an influencer posts a video ad in anticipation of product sales lift (e.g., if the product is being advertised on other channels). This concern may not apply in our setting because, as discussed, influencers are mainly motivated by engagement metrics instead of product sales. Moreover, according to our partner company, many sellers at the time of our study are small sellers who have limited advertising resources. Our practitioner interviews also suggest that influencers do not tend to choose ad posting time based on product-specific demand (Online Appendix D). Nevertheless, we collect the Baidu search index of each product in our sales panel as a proxy for its unobserved time-varying demand and include it as a control variable, $Search_{vd}$ (see Online

Appendix G for details).[38] Lastly, ϵ_{vd} is the error term.

Column (1) of Table 4 presents the ordinary least squares (OLS) regression result. Overall, posting an influencer video ad shows no significant effect on product sales. This insignificant result is worth noting given that sellers pay nontrivial amounts to advertise their products (influencers on average charge 19,530 RMB per video; see Table F.1). This result further highlights the importance of being able to predict sales lift before investing in an influencer video ad. Meanwhile, search intensity shows a positive and significant association with sales, as one would expect.

Table 4. Effect of Influencer Video Ads on Sales

	Sales Revenue					
	(1)	(2)	(3)	(4)	(5)	(6)
Post	−14.68	−78.02**	−73.75	18.14	8.10	−23.11
	(15.04)	(27.49)	(50.31)	(22.02)	(21.52)	(105.44)
Post × PE-Score		323.03**				636.10***
		(117.38)				(156.70)
Post × Engagement			125.45			−20.63
			(101.96)			(166.44)
Post × Product				−181.77*		−634.23
				(89.05)		(392.29)
Post × Engagement × Product					−265.90	292.25
					(179.60)	(823.19)
Search	7.37**	7.39**	7.39**	7.37**	7.36**	7.43**
	(2.76)	(2.76)	(2.76)	(2.76)	(2.76)	(2.76)
Post × Covariates	No	No	No	No	No	Yes
Video/Product, Influencer, Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	173,515	173,515	173,515	173,515	173,515	173,515
Adjusted R^2	0.98	0.98	0.98	0.98	0.98	0.98

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 (6) also controls for the interaction between Post and covariates including product characteristics (price and discount) and influencer characteristics (see Table F.1). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

In the second step to test our main hypothesis, we examine heterogeneous treatment

[38] The largest search engine in China, Baidu provides data on keyword-search dynamics, a service analogous to Google Trends. The Baidu Index has been used in academic research to control for unobserved market-level interest in various topics (e.g., Jia et al. 2020). For each product in our sales panel, we entered its brand on the Baidu Index platform to track its keyword-search records over the window of our data.

effects, of whether pe-score positively moderates the effect of influencer video ads on sales. We estimate the following specification:

$$\textit{Sales}_{vd} = \alpha \cdot \textit{Post}_{vd} + \beta \cdot \textit{Post}_{vd} \times \textit{PE-Score}_{v} + \textit{Video}_{v} + \textit{Influencer}_{i(v)} + \textit{Day}_{d} + \textit{Search}_{vd} + \epsilon_{vd}, \quad (3)$$

where $\textit{PE-Score}_{v}$ is the pe-score of video v. Note that the main effect of $\textit{PE-Score}_{v}$ cannot be separately identified from video/product fixed effects. Our main hypothesis is equivalent to the coefficient β being positive. As column (2) of Table 4 shows, this is indeed the case. Influencer video ads with higher pe-scores are significantly more effective in lifting sales.

To test whether it is product-related engagement that predicts sales lift, we also check whether overall engagement *alone* or product placement *alone* would have a similar moderation effect. As column (3) of Table 4 shows, the moderation effect of engagement, as measured by computed engagement to facilitate comparison, is statistically insignificant. Using observed engagement also yields an insignificant moderation effect. This result reaffirms industry observations and our correlational finding, that engagement does not necessarily predict sales lift.

Column (4) of Table 4 shows the moderation effect of product placement. It turns out that increasing product appearance can actually hurt sales, although the effect is less significant ($p < 0.05$). One possibility is that most users come to TikTok for entertainment, so unexpected and uninteresting promotions from influencers may damage brand perception via reactance (e.g., Edwards et al. 2002). This finding also echoes the "advertising avoidance" literature, where viewers dislike excessive ads when their viewing purpose is entertainment (e.g., Elpers et al. 2003, Wilbur 2016). Sellers and influencers may thus want to consider product placement as a limited budget; when, where, and how to present the product matters.

We also examine the effect of video ads that both engage the viewer and feature the product actively, but do so separately. For example, a video may be entertaining in the

first half and feature intensive product introduction in the second half, resulting in high engagement and high product scores but a low pe-score per our algorithm. To test the effect of these videos, we compute the interaction term between video-level computed engagement and video-level product placement as an alternative to pe-score. This term has an insignificant moderation effect on sales lift, as shown in column (5) of Table 4. Therefore, improving engagement and product placement separately may not help; it is important that these two are aligned spatiotemporally as captured by pe-score.[39]

Last, we include pe-score, engagement, product placement, and the engagement-product interaction term simultaneously as moderating variables for column (6) 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 F.1). We call this comprehensive model the "main specification." The moderation effect of pe-score remains positive and significant and is sizable. To put the effect magnitude in context, a 1% increase in pe-score is on average associated with a 6,361 RMB (about 920 USD) or a 2.58% increase in sales revenue of a product.

One thing to notice about Table 4 is the lack of variation in R^2 across specifications. We suspect that the reason has to do with the large number of fixed effects in all specifications, 2,684 for videos/products, 1,403 for influencers, and 166 for days. In practice, these factors may indeed account for substantial variation in sales, but may be costly to change. For example, the product line may take time to redesign, or the brand may have hired an influencer. Pe-score is intended as a metric for managing or selecting video ad content on top of these factors.

To more directly evaluate the predictive power of pe-score, we conduct a supplementary video-level analysis. We first compute each product's difference in average sales revenue before versus after the posting of its video ad, as in Figure 3. We then use XG-

[39]This result echos Zhang et al. (2020), who found that the temporal synchronicity between user-comment volume and movie content predicts movie enjoyment.

Boost[40] (Chen and Guestrin 2016) to predict this before-after sales difference with product features (category, price, discount), influencer characteristics (Table F.1) and video ad posting date. We train the prediction model on 80% of the videos in the sales sample and test it on the remaining 20%. Adding pe-score to the feature set reduces the root mean squared error (RMSE) on the test set from 2,937 to 2,549, a 13% reduction. This video-level analysis provides further evidence that pe-score is informative of ad effectiveness.

We further conduct extensive robustness analyses. Online Appendix I presents the details. To summarize, our results are robust with respect to alternative construction (I.1) and validation (I.2) of the engagement heatmap, and various causal identification strategies including pre-trend and dynamic-effect testing (I.3), continuous-treatment-effect estimation (I.4), simulation of different treatment specifications (I.5), randomization inference (I.6), and alternative diff-in-diff specifications (I.7).

5.4 PE-Score Predicts Better for Low-Involvement Products

In this section, we extend our main analysis to investigate whether pe-score predicts sales lift better in some product categories than others. This will help us understand where and why our algorithm works.

To characterize the 11 product categories in our sample, we randomly surveyed 175 TikTok users on Wenjuanxing, a major survey platform in China. These users answered yes to our screening question: have you watched any ads on TikTok? We asked 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 were similar between our survey sample and users on TikTok around the time of the survey.[41]

[40] https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf.
[41] Douyin user profile – research report. *Industry research report*, April 30, 2020.

Figure H.1 of the Online Appendix summarizes the average ratings across product categories. 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 pe-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 price level which we observe from our data. (We do not introduce the median indicators of these variables as interaction terms in a pooled regression because the existing model contains three-way interaction already.) We then re-estimate our main specification on the split subsamples. The estimation results reveal a noticeable pattern. Pe-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.[42] The result for informativeness depends on where the equal sign goes, suggesting that sample size may be the main driver of statistical significance.

These results are consistent with the buying process underlying the pe-score concept: users come to the platform for entertainment; an engaging ad may activate their attention, which translates into purchase if buying does not require much cognitive processing. In other words, a higher pe-score is more likely to translate into higher sales 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, or what we summarize as low-involvement categories. These categories commonly appear on entertainment commerce platforms; they together represent over 90% of products in our TikTok data (see Figure F.4 of the Online Appendix).

[42] The conclusion is also robust if we perform median splits at the product level.

Table 5. Predictive Power of PE-Score by Product Category

	Sales Revenue					
	Deliberate ≤ Median	Deliberate > Median	Hedonic ≤ Median	Hedonic > Median	Price ≤ Median	Price > Median
Post	4.26	25.77	21.11	4.59	365.00***	−114.08
	(142.85)	(156.31)	(153.84)	(143.30)	(52.14)	(232.34)
Post × PE-Score	800.45***	−151.62	−183.48	803.60***	987.15***	450.51
	(183.52)	(287.18)	(269.98)	(185.81)	(64.32)	(419.19)
Post × Engagement	−53.68	−73.90	−53.03	−54.61	−454.47***	24.71
	(191.70)	(327.67)	(320.49)	(192.30)	(79.97)	(404.56)
Post × Product	−844.20	6.12	36.14	−844.41	−4,152.79***	320.76
	(445.83)	(775.24)	(762.37)	(447.21)	(235.78)	(733.46)
Post × Engagement × Product	571.76	688.44	602.74	569.17	5,753.93***	−921.93
	(944.38)	(1,560.53)	(1,534.53)	(947.35)	(460.64)	(1,547.59)
Search	6.61	8.73**	8.76**	6.60	−0.24	8.89*
	(3.78)	(2.88)	(2.84)	(3.80)	(1.97)	(4.37)
Post × Covariates	Yes	Yes	Yes	Yes	Yes	Yes
Video/Product, Influencer, Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	143,151	30,364	31,209	142,306	90,859	82,656
Adjusted R^2	0.98	0.71	0.72	0.98	0.79	0.98

Note: The full sample is the sales panel, where each product has one video ad. Each observation is at the product-day level. The split on deliberate and hedonic is based on the median of average ratings in the survey across 11 product categories. The split on price is based on the median of all products in the sales panel. 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 F.1). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Notably, video as a popular content format of entertainment commerce may be more effective at influencing low-involvement purchase decisions compared with print media (Liu et al. 2020). As such, we expect pe-score to be a particularly suitable predictor of sales conversion in entertainment commerce.

Extending the cross-category analysis, we find that pe-score is significantly higher in video ads for influencers' own products than others' products in impulsive and hedonic categories (Table 6). In other words, incentive misalignment between the influencer and the seller is more of a problem in categories where pe-score matters more. This result is intuitively appealing because, if pe-score does not matter, the influencer might as well focus on being engaging. The exception is that we do not find a significant difference in pe-score by price level. A possible explanation is that although pe-score is more effective for inexpensive products, lower prices reduce the monetary return from product sales

when influencers advertise their own products, which shifts influencers' focus towards engagement even if they are advertising their own products.

Table 6. PE-Score by Influencer Advertising Own versus Others' Products

	PE-Score			
	(1)	(2)	(3)	(4)
Own	26.50**	36.34***	−10.68	23.43
	(9.85)	(10.52)	(16.92)	(13.47)
Deliberate > Median		23.39		
		(12.65)		
Hedonic > Median			−21.32	
			(12.26)	
Price > Median				0.61
				(2.57)
Own × Deliberate > Median		−53.17**		
		(20.14)		
Own × Hedonic > Median			48.69**	
			(18.02)	
Own× Price > Median				5.15
				(15.41)
Covariates	Yes	Yes	Yes	Yes
Observations	2,762	2,762	2,762	2,762
Adjusted R^2	0.02	0.02	0.02	0.02

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. Controlling for product price, discount, and category and influencer characteristics (see Table F.1). $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

6 Exploring Drivers of PE-Score

We have seen that pe-score varies across videos and this variation matters in predicting sales lift. But ideally, we want to go beyond predictive analysis to offer prescriptive insight on what substantive measures an influencer can take to improve pe-score. Recall that pe-score captures the spatiotemporal synchronicity between the engagement and product heatmaps. The product heatmap has a straightforward interpretation; it captures product presence. The engagement heatmap is less interpretable; it outputs the more engaging regions of a video without offering an explanation. We explore this issue next.

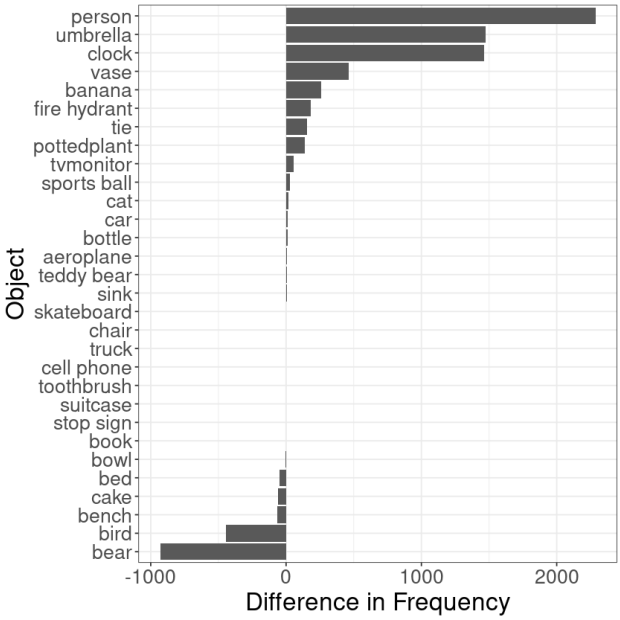
An established approach would be to use proven theories to guide the interpretation of unstructured data. For example, Zhang et al. (2020) used film grammar to analyze movie content and Zhang et al. (2021) used photography theory to evaluate image quality. The challenge in our setting, as confirmed by our partner company, is that there is not yet a widely accepted theory on how to make influencer video ads engaging. Therefore, we explore drivers of engagement in a bottom-up, data-driven way. We do so at three levels – pixel, frame, and video segment – motivated by a natural sequence of questioning.

Given the engagement heatmap, the first question is what objects tend to appear in high engagement pixels. To answer this question, we divide the pixels in a video into high and low types based on a median split on pixel-level engagement scores. Then we create two versions of the video: one that only uses high engagement pixels with low engagement pixels blacked out (high version), and the reverse (low version). We then run an object detection algorithm, YOLO (Redmon et al. 2016),[43] on the high and low versions of the same video to identify what objects, from 80 pre-specified classes, are presented in each version. For each detected object, we compute its net frequency of appearance in high versus low versions of all videos in the sales panel. Figure 4 presents the results. More object instances are detected in high engagement pixels (9,916) than low ones (4,780). Moreover, as one would expect, humans are the most represented class in high engagement pixels.

Based on the finding that human presence is a key part of engagement, the next question is what humans can do in the video to engage. 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). Indeed, as a sanity check of our algorithm, we find a positive and significant correlation between the presence of human faces and pixel-level engagement (Online Appendix I.2). We further ask what facial expressions drive engagement, for two reasons. First, facial expressions are arguably more actionable than factors such

[43] https://pjreddie.com/darknet/yolo.

Figure 4. Objects in High versus Low Engagement Pixels



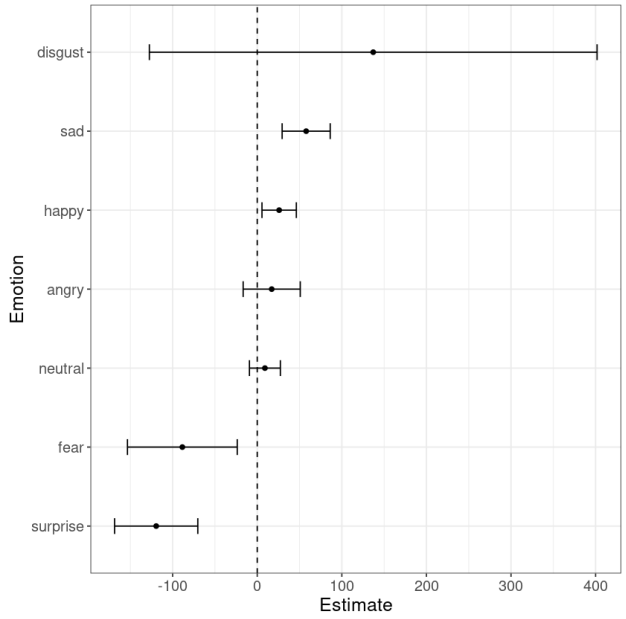
Note: Frequency is the number of times an object is detected in either the high or low version of videos in the sales panel. Difference in frequency is the frequency in high versions minus that in low versions. Some objects are only detected in high or low versions alone so that a difference cannot be computed, but these objects are rare.

as facial attractiveness. Second, given the prevalence of low-involvement categories on TikTok, engaging users emotionally may be an effective route to conversion. We run an emotion detection algorithm, FER (Zhang et al. 2016, Arriaga et al. 2017),[44] that detects facial expressions of Ekman (1992)'s six basic emotions (anger, disgust, fear, happiness, sadness, and surprise) plus a neutral emotion. We apply FER to each frame of each video in the sales panel. Then we regress the average pixel-level engagement scores on a given frame on the detected emotions. The results are shown in Figure 5. Happiness and sadness are positively and significantly associated with engagement scores, whereas fear and surprise are negatively and significantly associated. The pattern echoes earlier findings in the literature. For example, Wild et al. (2001) found that happiness and sadness are particularly contagious, taking effect in as short as half a second.

Lastly, to expand the space of actionable recommendations beyond facial expressions,

[44]https://pypi.org/project/fer.

Figure 5. Emotions and Engagement



Note: Each observation is a frame of a video in the sales panel. Results are relative to a baseline where no emotion is detected. Bars are the 95% confidence intervals.

we ask what actions an influencer can take in the video to engage the audience. For a broad search of possible actions, we again take the data-driven approach by detecting what activities are in the video and how they relate to engagement. We divide each video in the sales panel into one or more segments that each lasts 15 seconds. We run an activity detection algorithm, I3D (Carreira and Zisserman 2017),[45] on each segment to classify it into 400 pre-specified classes.[46] Then we regress the average pixel-level engagement score in a segment on the detected activity controlling for segment sequence (e.g., the second in a video). We find 112 activities that are positively associated with engagement scores and 40 activities that are negatively associated. The top 30 activities by effect size that are significantly associated with engagement scores ($p < 0.05$) are reported in Figure J.1 of Online Appendix J. We can see that positive activities tend to be relatively more energetic and fast-paced (e.g., side kick, salsa dancing, krumping) or novel (e.g., getting a tattoo,

[45] https://github.com/deepmind/kinetics-i3d.

[46] Shorter segments may capture finer dynamics in a video but give the algorithm less data to work with in each segment. We also tried segments of 5 or 10 seconds but the algorithm did not reliably identify activities in these shorter segments.

snorkeling, contact juggling), whereas negative activities tend to be relatively slow-paced (e.g., swimming backstroke, knitting, playing monopoly) or mundane (e.g., reading book, setting table, stretching arm).

We take two approaches to validate this hypothesis. First, we use topic modeling to uncover any underlying themes among these activities. Table J.1 in Online Appendix J lists the top 10 words in two topics for activities positively or negatively associated with engagement, respectively.[47] Indeed, words with higher energy, pace, or novelty (e.g., ski, kick, dance, blow, climb, basketball) are identified in more-engaging activities. Words with lower energy, pace, or are more mundane (e.g., wax, answer, paper, feed, share, trim) are identified in less-engaging activities.

Second, we conduct another survey with 104 college students and staff members to identify the commonality in more versus less-engaging activities. These participants tend to be familiar with TikTok. As such, they may be able to interpret these activities in the context of TikTok beyond what topic modeling can reveal. Each participant was asked to write three to five adjectives or phrases to indicate their perception of the common characteristics. We plot the word clouds based on their responses for more versus less-engaging activities in Figure 6. Participants tend to use words such as interesting, novel, funny, skill, and stimulating to describe more-engaging activities, whereas for less-engaging activities, they tend to use words such as boring/uninteresting, ordinary/common/routine, simple, dull, and slow pace.

Combining results from the analyses at pixel, frame, and video-segment levels, we find that human presence, sad or happy emotions, and stimulating or novel activities are positively associated with engagement. To improve pe-score, it may be helpful to spatiotemporally align product placement with these elements of engagement. For example, it may be helpful to feature a product in the moment of high emotional connectivity, or with an influencer who is performing a stimulating activity.

[47] The optimal number of topics is selected via Cao et al. (2009) and Deveaud et al. (2014).

Figure 6. Word Clouds from Survey of Perceived Activity Characteristics

More-Engaging Activities Less-Engaging Activities

7 Concluding Remarks

In this paper, we propose an algorithm to compute a summary statistic called pe-score, which draws on unstructured video data to predict the effect of influencer video ads on product sales. This summary statistic computes a product engagement score 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 20,000 TikTok influencer video ads and their corresponding product sales. In the data, raw 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 via pe-score and show that video ads with higher pe-scores are significantly more effective in lifting sales. The effect is sizable. A 1% increase in pe-score is on average associated with a \$920 or 2.58% increase in sales revenue of a product.

In further analysis, we find that pe-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 pe-score. Lastly, we find that engagement increases with human presence, sad or happy emotions, and stimulating or novel activities.

Pe-score can potentially unlock a myriad of applications. A practical advantage of pe-score is that it can be computed based on our trained algorithm 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 scalable and can be used to evaluate a large number of candidate videos quickly. Our algorithm is also applicable beyond the placement of physical products in a video ad. The product can be replaced by a brand name, logo, or any key message that needs to be conveyed, as long as the message is visually detectable in the video. Various stakeholders in the influencer advertising space can potentially benefit from these features.

More specifically, influencers can use pe-score to aid video content development. They can make a video more engaging using the actionable drivers behind pe-score, place the product in the engaging pixels, and check the resulting pe-score for real-time feedback. Sellers can use pe-score as a novel contractual instrument. For example, sellers can compensate influencers based on the pe-score of their video ads. In comparison, the current industry norm of engagement-based compensation may exacerbate incentive misalignment, 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, pe-score can serve as a metric to help clarify the attribution of sales outcome between sellers and influencers. Finally, entertainment commerce platforms can leverage pe-score to launch various features to improve transaction efficiency. For example, a platform can highlight pe-score as a key performance index of influencers. Providing pe-score alongside engagement metrics can help sellers choose influencers and manage campaigns with richer information.

There are several directions for future research. First, it will be interesting to study various applications of the algorithm and track their impact on the entertainment commerce industry. Second, pe-score is learned mainly through the visual component of a video ad while controlling for the acoustic features and spoken content. How to better integrate the two in an interpretable way is a worthy question. Third, our exploration of engagement drivers is correlational and preliminary. Controlled experimentation may unlock further insight. Lastly, it will be meaningful to explore the generalizability of pe-score. We validated the algorithm in the context of influencer video ads, where pe-score fundamentally matters because it captures the importance of attention in entertainment commerce and because influencers may strategically choose pe-score as a result of incentive misalignment. However, the general principle of making product placement engaging should extend to other forms of video advertising. It will be encouraging if the algorithm is able to predict sales lift based on the mere content of a generic video ad.

References

Arriaga, O., M. Valdenegro-Toro, and P. Plöger (2017). Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*.

Avery, J. and A. Israeli (2020). Influencer marketing. *Harvard Business School Case*, N9–520–075.

Azer, J. and M. Alexander (2020). Negative customer engagement behaviour: the interplay of intensity and valence in online networks. *Journal of Marketing Management 36*(3-4), 361–383.

Bakhshi, S., D. A. Shamma, and E. Gilbert (2014). Faces engage us: photos with faces attract more likes and comments on Instagram. *CHI '14: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 965–974.

Barnett, S. B. and M. Cerf (2017). A ticket for your thoughts: method for predicting content recall and sales using neural similarity of moviegoers. *Journal of Consumer Research 44*(1), 160–181.

Baumgartner, H., M. Sujan, and D. Padgett (1997). Patterns of affective reactions to advertisements: the integration of moment-to-moment responses into overall judgments. *Journal of Marketing Research 34*(2), 219–232.

Bertrand, M., E. Duflo, and S. Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics 119*(1), 249–275.

Borji, A., H. R. Tavakoli, D. N. Sihite, and L. Itti (2013). Analysis of scores, datasets, and models in visual saliency prediction. *Proceedings of the IEEE international conference on computer vision*, 921–928.

Borusyak, K., X. Jaravel, and J. Spiess (2022). Revisiting event study designs: Robust and efficient estimation.

Brown, D. and N. Hayes (2008). *Influencer marketing*. Routledge.

Burnap, A., J. R. Hauser, and A. Timoshenko (2021). Design and evaluation of product aesthetics: a human-machine hybrid approach. *SSRN 3421771*.

Callaway, B. and P. H. Sant'Anna (2021). Difference-in-differences with multiple time periods. *Journal of Econometrics 225*(2), 200–230.

Cao, J., T. Xia, J. Li, Y. Zhang, and S. Tang (2009). A density-based method for adaptive lda model selection. *Neurocomputing 72*(7-9), 1775–1781.

Carreira, J. and A. Zisserman (2017). Quo vadis, action recognition? a new model and the kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6299–6308.

Chaturvedi, I., K. Thapa, S. Cavallari, E. Cambria, and R. E. Welsch (2021). Predicting video engagement using heterogeneous deepwalk. *Neurocomputing 465*, 228–237.

Chen, T. and C. Guestrin (2016). XGBoost: a scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 785–794.

Cheng, M. M. and S. Zhang (2022). Reputation burning: analyzing the impact of brand sponsorship on social influencers. *SSRN 4071188*.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1251–1258.

De Chaisemartin, C. and X. d'Haultfoeuille (2020). Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review 110*(9), 2964–96.

Deveaud, R., E. SanJuan, and P. Bellot (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique 17*(1), 61–84.

Dew, R., A. Ansari, and O. Toubia (2022). Letting logos speak: Leveraging multiview representation learning for data-driven branding and logo design. *Marketing Science 41*(2), 401–425.

Du, R., Y. Zhong, H. Nair, B. Cui, and R. Shou (2019). Causally driven incremental multi touch attribution using a recurrent neural network. *AdKDD Workshop, 2019 KDD Conference, Anchorage*.

Dupont, L., K. Ooms, M. Antrop, and V. Van Eetvelde (2016). Comparing saliency maps and eye-tracking focus maps: the potential use in visual impact assessment based on landscape photographs. *Landscape and Urban Planning 148*, 17–26.

Edwards, S. M., H. Li, and J.-H. Lee (2002). Forced exposure and psychological reactance: Antecedents and consequences of the perceived intrusiveness of pop-up ads. *Journal of advertising 31*(3), 83–95.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion 6*(3-4), 169–200.

Elpers, J. L. W., M. Wedel, and R. G. Pieters (2003). Why do consumers stop viewing television commercials? Two experiments on the influence of moment-to-moment entertainment and information value. *Journal of Marketing Research 40*(4), 437–453.

Gardner, J. (2022). Two-stage differences in differences. *arXiv preprint arXiv:2207.05943*.

Goodman-Bacon, A. (2018). Difference-in-differences with variation in treatment timing. Technical report, National Bureau of Economic Research.

Hartmann, J., M. Heitmann, C. Schamp, and O. Netzer (2021). The power of brand selfies. *Journal of Marketing Research 58*(6), 1159–1177.

Hou, X. and L. Zhang (2007). Saliency detection: a spectral residual approach. *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Itti, L. (2005). Models of bottom-up attention and saliency. *Neurobiology of attention*, 576–582.

Jakiela, P. (2021). Simple diagnostics for two-way fixed effects. *arXiv preprint arXiv:2103.13229*.

Jia, J. S., X. Lu, Y. Yuan, G. Xu, J. Jia, and N. A. Christakis (2020). Population flow drives spatio-temporal distribution of COVID-19 in China. *Nature 582*(7812), 389–394.

John, L. K., O. Emrich, S. Gupta, and M. I. Norton (2017). Does "liking" lead to loving? The impact of joining a brand's social network on marketing outcomes. *Journal of Marketing Research 54*(1), 144–155.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Lee, D., K. Hosanagar, and H. S. Nair (2018). Advertising content and consumer engagement on social media: evidence from Facebook. *Management Science 64*(11), 5105–5131.

Levitt, S. D. and C. Syverson (2008). Market distortions when agents are better informed: the value of information in real estate transactions. *Review of Economics and Statistics 90*(4), 599–611.

Li, X., M. Shi, and X. S. Wang (2019). Video mining: measuring visual information using automatic methods. *International Journal of Research in Marketing 36*(2), 216–231.

Li, Y. and Y. Xie (2020). Is a picture worth a thousand words? An empirical study of image content and social media engagement. *Journal of Marketing Research 57*(1), 1–19.

Little, J. D. C. (1979). Aggregate advertising models: the state of the art. *Operations research 27*(4), 629–667.

Liu, L., D. Dzyabura, and N. Mizik (2020). Visual listening in: extracting brand image portrayed on social media. *Marketing Science 39*(4), 669–686.

Liu, Q., H. Liu, and M. Kalwani (2020). "See your doctor": the impact of direct-to-consumer advertising on patients with different affliction levels. *Marketing Letters 31*(1), 37–48.

Liu, X., S. W. Shi, T. Teixeira, and M. Wedel (2018). Video content marketing: the making of clips. *Journal of Marketing 82*(4), 86–101.

Lou, C. and S. Yuan (2019). Influencer marketing: how message value and credibility affect consumer trust of branded content on social media. *Journal of Interactive Advertising 19*(1), 58–73.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the seventh IEEE international conference on computer vision 2*, 1150–1157.

Malik, N. and P. V. Singh (2019). Deep learning in computer vision: methods, interpretation, causation, and fairness. *Operations Research & Management Science in the Age of Analytics*, 73–100.

Mitchell, A. A. and J. C. Olson (1981). Are product attribute beliefs the only mediator of advertising effects on brand attitude? *Journal of Marketing Research 18*(3), 318–332.

Pei, A. and D. Mayzlin (2022). Influencing social media influencers through affiliation. *Marketing Science 41*(3), 593–615.

Polsfuss, M. and M. Hess (1991). Liking through moment-to-moment evaluation; identifying key selling segments in advertising. *Advances in Consumer Research 18*, 540–544.

Rajaram, P. and P. Manchanda (2020). Video influencers: unboxing the mystique. *SSRN 3752107*.

Ramanathan, S. and G. Menon (2006). Time-varying effects of chronic hedonic goals on impulsive behavior. *Journal of Marketing Research 43*(4), 628–641.

Redmon, J., S. Divvala, R. Girshick, and A. Farhadi (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.

Sala-i Martin, X. X. (1997). I just ran four million regressions. Technical report, National Bureau of Economic Research.

Salman, S. and X. Liu (2019). Overfitting mechanism and avoidance in deep neural networks. *arXiv preprint arXiv:1901.06566*.

Schouten, A. P., L. Janssen, and M. Verspaget (2020). Celebrity vs. influencer endorsements in advertising: the role of identification, credibility, and product-endorser fit. *International Journal of Advertising 39*(2), 258–281.

Shapiro, B. T., G. J. Hitsch, and A. E. Tuchman (2021). TV advertising effectiveness and profitability: generalizable results from 288 brands. *Econometrica 89*(4), 1855–1879.

Simonyan, K., A. Vedaldi, and A. Zisserman (2013). Deep inside convolutional networks: visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 1929–1958.

Sun, L. and S. Abraham (2020). Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *arXiv preprint arXiv:1804.05785*.

Teixeira, T., R. Picard, and R. El Kaliouby (2014). Why, when, and how much to entertain consumers in advertisements? A web-based facial tracking field study. *Marketing Science 33*(6), 809–827.

Teixeira, T. S., M. Wedel, and R. Pieters (2010). Moment-to-moment optimal branding in TV commercials: preventing avoidance by pulsing. *Marketing Science 29*(5), 783–804.

Testwuide, T. (2020). Why marketing attribution has failed in the boardroom. *Forbes*.

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

Tong, L. C., M. Y. Acikalin, A. Genevsky, B. Shiv, and B. Knutson (2020). Brain activity forecasts video engagement in an internet attention market. *Proceedings of the National Academy of Sciences 117*(12), 6936–6941.

Tran, D., L. Bourdev, R. Fergus, L. Torresani, and M. Paluri (2015). Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, 4489–4497.

Tucker, C. E. (2015). The reach and persuasiveness of viral video ads. *Marketing Science 34*(2), 281–296.

Wedel, M. and R. Pieters (2008). *Eye Tracking for Visual Marketing*. Now Publishers Inc.

Wernerfelt, B., A. J. Silk, and S. Yu (2021). Internalization of advertising services: Testing a theory of the firm. *Marketing Science 40*(5), 946–963.

Wilbur, K. C. (2016). Advertising content and television advertising avoidance. *Journal of Media Economics 29*(2), 51–72.

Wild, B., M. Erb, and M. Bartels (2001). Are emotions contagious? Evoked emotions while viewing emotionally expressive faces: quality, quantity, time course and gender differences. *Psychiatry Research 102*(2), 109–124.

Zhang, C., S. Bengio, M. Hardt, B. Recht, and O. Vinyals (2017). Understanding deep learning requires rethinking generalization. *5th International Conference on Learning Representations*.

Zhang, K., Z. Zhang, Z. Li, and Y. Qiao (2016). Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters 23*(10), 1499–1503.

Zhang, M. and L. Luo (2022). Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp. *Management Science*, forthcoming.

Zhang, Q., W. Wang, and Y. Chen (2020). Frontiers: in-consumption social listening with moment-to-moment unstructured data: the case of movie appreciation and live comments. *Marketing Science 39*(2), 285–295.

Zhang, Q. and S. Zhu (2018). Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering 19*(1), 27–39.

Zhang, S., D. Lee, P. V. Singh, and K. Srinivasan (2021). What makes a good image? Airbnb demand analytics leveraging interpretable image features. *Management Science*, forthcoming.

Online Appendix

A 3D Convolutional Neural Network (3D CNN)

We use a 3D CNN and gradient-based saliency map to estimate the 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 (S, 224, 224, 3) numerical array, where S 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.

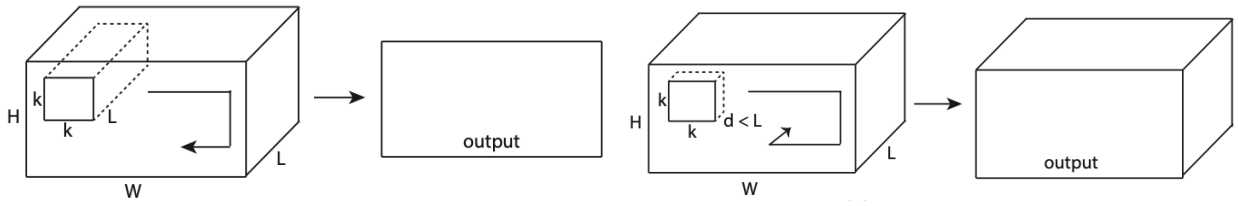
The key building blocks of a CNN are convolution layers. A convolution layer uses filters with weights that are trainable to transform the input images by representing them in a more abstract feature space that captures more general properties of the images (e.g., the presence of an edge or face). What 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.

CNN has been used to analyze images for marketing research in a few recent papers (e.g., Liu et al. 2020, Tkachenko and Jedidi 2020, Hartmann et al. 2021, Zhang et al. 2021, Zhang and Luo 2022). These papers are built upon 2D CNN. 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 and the input always have the same depth L, which represents the three color channels. The filter 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 $d < L$, where L represents the three color channels and time. In addition to sliding across the spatial dimensions, the filter also slides across the depth dimension, outputting a 3D array.

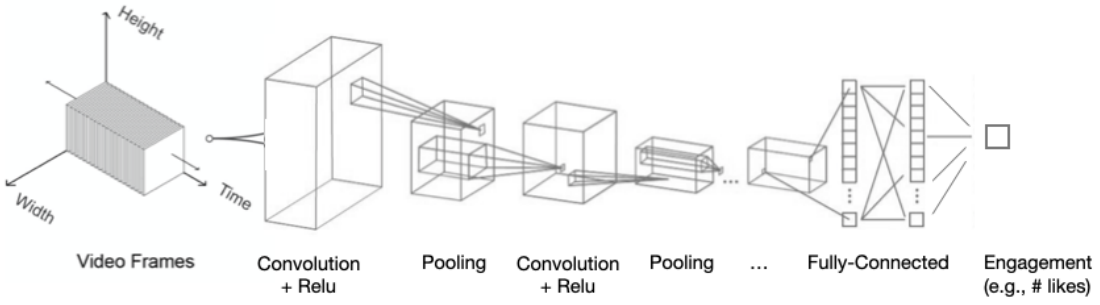
Figure A.1. 2D versus 3D Convolution (Tran et al. 2015)



2D Convolution 3D Convolution

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

Figure A.2. A Stylized 3D CNN Architecture for Engagement Prediction



More specifically, we build on Xception (Chollet 2017) pre-trained on ImageNet to extract features from each frame (in a time-distributed manner). Because the top layer of

[1] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

Xception performs a classification task, we remove the top layer while keeping weights in other layers frozen. We stack a 3D convolution layer with 128 units and a filter size of (3, 3, 3) 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. The standard max pooling layer outputs a feature map whose dimension depends on the dimension of the input feature map produced by the 3D convolution layer. This approach does not work in our case because our algorithm takes in videos of different lengths. We instead use a global max pooling layer to map variable input feature map dimensions into a fixed dimension. We then add a 128-unit dense layer, and a dropout layer (with a dropout rate of 0.1) which has been shown to be particularly effective at reducing overfitting (Srivastava et al. 2014) on top of the global max pooling layer. The final layer is a one-unit dense layer to output the predicted engagement of a video.

The model is optimized with Adadelta[2] against the mean absolute percentage error (MAPE) loss with an initial learning rate of 0.001 that is adjusted adaptively in the training process. The architecture of our network on top of Xception is summarized in Figure A.3.[3] Hyperparameters such as the number of units in the 3D convolution layer and dense layer, filter size, dropout rate, and initial learning rate are tuned via a grid search on a smaller training sample with 1,000 videos. The optimal combination is chosen by validation error as detailed below.

As explained in the paper, prior to training, we regress raw engagement on product fixed effects, influencer fixed effects, acoustic features, and spoken content. We then normalize the regression residuals to $[0, 1]$ for training. To derive acoustic features, we 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 for tractability. To derive text features, We extract the transcript (mostly

[2]https://keras.io/api/optimizers/adadelta.

[3]Xception has 132 layers, hence our full network has 132 + 6 = 138 layers. See Chollet (2017) for more details on the architecture of Xception.

Figure A.3. CNN Layers on Top of Xception

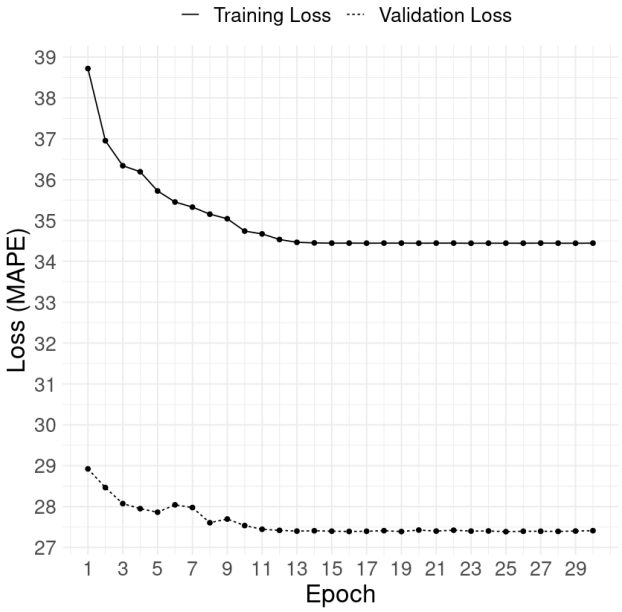
```
 Layer (type)                 Output Shape              Param #
=================================================================
 time_distributed_3 (TimeDis  (None, None, 7, 7, 2048)  20861480
 tributed)

 conv3d_3 (Conv3D)            (None, None, 5, 5, 128)   7078016

 global_max_pooling3d (Globa  (None, 128)               0
 lMaxPooling3D)

 dense_6 (Dense)              (None, 256)               33024

 dropout_3 (Dropout)          (None, 256)               0

 dense_7 (Dense)              (None, 1)                 257

=================================================================
Total params: 27,972,777
Trainable params: 7,111,297
Non-trainable params: 20,861,480
_________________________________________________________________
```

in Chinese) of each video using Google Speech-to-Text API[4]. We then use a pre-trained multilingual BERT model[5] to convert the transcripts into 768-dimension embeddings.

Our final 3D CNN has more than 7 million trainable parameters. We train it on 10,000 videos and validate it on 3,500 videos starting from weights optimized on 1,000 videos to speed up convergence. The model is trained with GPUs on a high performance computing cluster using TensorFlow (https://keras.io). Figure A.4 summarizes the training and validation loss statistics over 30 epochs. The training losses are higher than validation losses because a dropout layer is used in training but not in validation. Both loss curves become flat as the number of epochs increases and do not suggest signs of overfitting (e.g., Salman and Liu 2019). We retain the parameters at the epoch with the minimal validation error as the final model (epoch 25). The accuracy (one minus MAPE) on the holdout test set of 3,451 videos is 73%, which is comparable with recent results on predicting video ad engagement. For example, Chaturvedi et al. (2021) predicted watch time on YouTube video ads with a graph-embedding model and reported an accuracy of 78%.

[4]https://codelabs.developers.google.com/codelabs/cloud-speech-text-python3.
[5]https://tfhub.dev/tensorflow/bert_multi_cased_L-12_H-768_A-12.

Figure A.4. Training and Validation Loss



Note: Minimal validation loss is achieved at epoch 25. The training losses are higher than validation losses because a dropout layer is used in training but not validation.

As our main contribution is the concept of pe-score, not a new predictive model that achieves higher accuracy; any model that can be used to generate saliency maps (such as CNN or transformer-based models) can be implemented in our framework. We used one of the state-of-the-art models (3D CNN) as a proof of concept and are open to the possibility that other predictive models may enhance pe-score's efficacy in the future.

B Engagement Heatmap

We compute the engagement heatmap as a saliency map. A saliency map is a gradient-based visualization method for CNN (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 images with color, 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 entire video. This allows us to capture any dependency across video frames when deciding which pixels are driving engagement. More formally, we define pixel-level engagement as:

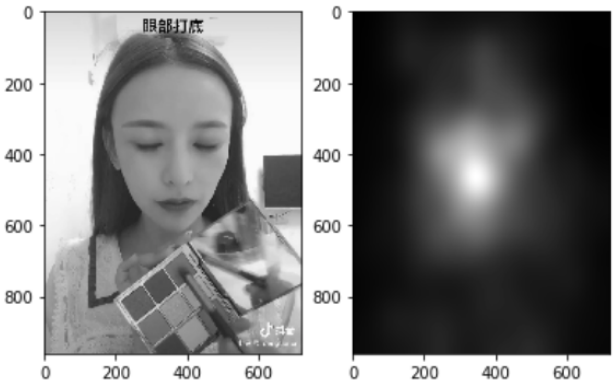
$$e_{hws} := \max_{\{r,g,b\}} \left(\left| \frac{\partial \hat{f}}{\partial x_{hwsr}} \right|, \left| \frac{\partial \hat{f}}{\partial x_{hwsg}} \right|, \left| \frac{\partial \hat{f}}{\partial x_{hwsb}} \right| \right)$$

where $\hat{f}$ is the trained 3D CNN, and x_{hwsr}, x_{hwsg}, and x_{hwsb} are the pixel values in the three color channels, respectively, 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).

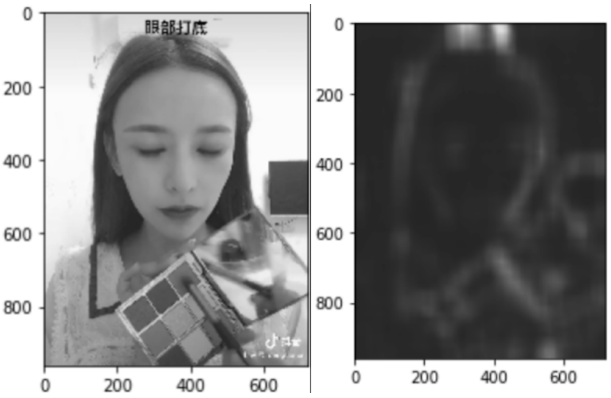
We use a saliency map to compute pixel-level engagement and call it the (supervised) engagement heatmap. It is supervised because the saliency map builds on a 3D CNN trained on video-level engagement data. In Online Appendix I.1, we also discuss an unsupervised approach to engagement heatmap that only requires the video itself.

We implement the supervised saliency map with tf-keras-vis[6] and the unsupervised saliency map with the saliency module in OpenCV.[7] Figure B.1a presents an example of a video frame, and its corresponding frame in the supervised engagement heatmap. Figure B.1b presents an example of an unsupervised engagement heatmap.

Figure B.1. An Example of the Engagement Heatmap



(a) Supervised Engagement Heatmap



(b) Unsupervised Engagement Heatmap

Note: The 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 engagement heatmap is in the right column (supervised on the top, unsupervised at the bottom). Brighter areas in the engagement heatmap correspond to pixels with higher saliency.

[6]https://github.com/keisen/tf-keras-vis.
[7]https://docs.opencv.org/master/d8/d65/group.html.

C Product Heatmap

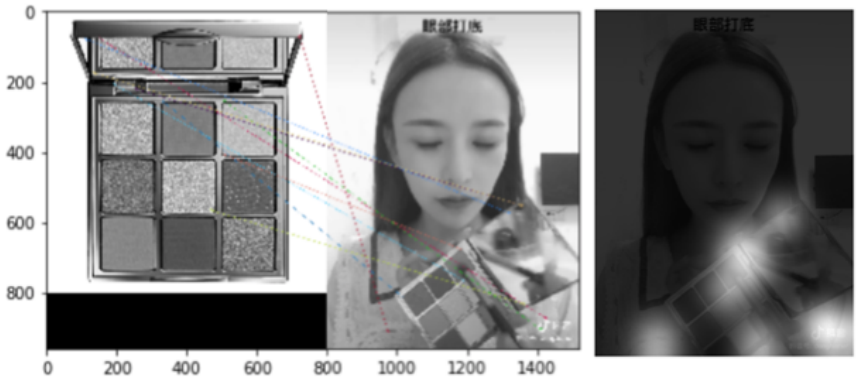
We use SIFT to detect whether an advertised product appears in a given pixel of the video. We implement SIFT via Oriented FAST and Rotated BRIEF (ORB) in OpenCV.[8]

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 product 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 a corresponding frame from the product heatmap of the video. The bright areas correspond to pixels where SIFT detects product presence. We can also see that the detected product pixels can be scattered in the frame and do not necessarily enclose the entire product. We, therefore, create a convex hull of the detected product pixels and consider all pixels within the hull as product pixels.

[8] https://docs.opencv.org/3.4/d1/d89/tutorial_py_orb.html.

Figure C.1. An Example of the Product Heatmap



Note: We use SIFT to detect the product (left column) in a video frame (middle column). The corresponding frame in the 3D product heatmap of the video is shown in the right column, where the bright areas indicate product presence. Random noises are added around detected product pixels to aid visualization.

D Transcript of Practitioner Interviews

To better understand the institutional background of influencer video advertising on TikTok, we interviewed a number of practitioners in the space. We present the interview transcript below (translated into English).

Interviewee: ThinkCrow, TikTok Influencer with 1.6 million followers in the science book (lifestyle) category.

- Question: Who determines the content design of the video?
  Answer: We have a content design team responsible for this.
- Question: Will advertisers interfere with content design?
  Answer: No interference at all.
- Question: How do you determine when to post an advertising video?
  Answer: The advertising time does not affect the result very much. There is no special design.

Interviewee: Yuerong Zhao, Senior Project Manager of a Tiktok influencer incubation company, which has more than 30 influencers in the household and makeup categories.

- Question: Who determines the content design of the video?
  Answer: Our content design team.
- Question: Will advertisers interfere with content design?
  Answer: It depends on the strength and popularity of the advertiser (brand). Powerful brands may interfere a little, but not too much. They will not interfere with the position of the products in the video.
- Question: How do you determine when to post an advertising video?
  Answer: There is not much planning for the ad posting time. Sometimes, ads will be posted before the Double 11 Festival. In most cases, there is no specific time.

- Question: What other advertising channels do your customers (advertisers) have?
  Answer: We have an exclusive agreement not to release the ad elsewhere.

Interviewee: Name undisclosed, the person in charge of Tiktok e-commerce live broadcast products in all categories.

- Question: Will advertisers interfere with content design?
  Answer: Sometimes the influencer is asked to speak for a specific amount of time. Sometimes there may be materials suggested for the influencer to use.

Interviewee: Jian Qin, Senior Product Manager of TikTok

- Question: Will advertisers interfere with content design?
  Answer: Generally, influencer companies have scripts, or video samples taken in the past. Some advertisers will also provide the advertising language and scripts (all text) they want to display.

Interviewee: Lei Zhou, Xingtu Advertising business affiliate.

- Question: Will advertisers interfere with content design?
  Answer: If the advertiser's company is very small and they want to spend less money, they won't care much about how the content is designed. If it is a large business, the advertiser will review the video to make sure there is no text content that damages the brand image. However, advertisers generally do not interfere with the location and time of the product placement and the video production method. Strong influencers are hardly interfered by advertisers.

In summary, these interviews suggest that, indeed, (1) sellers do not tend to influence the visual aspect of video content that we focus on in the paper, (2) sellers do not tend to influence product placement, and (3) influencers do not tend to choose the posting time of video ads based on product-specific demand.

E Predicting Missing Product Category Information

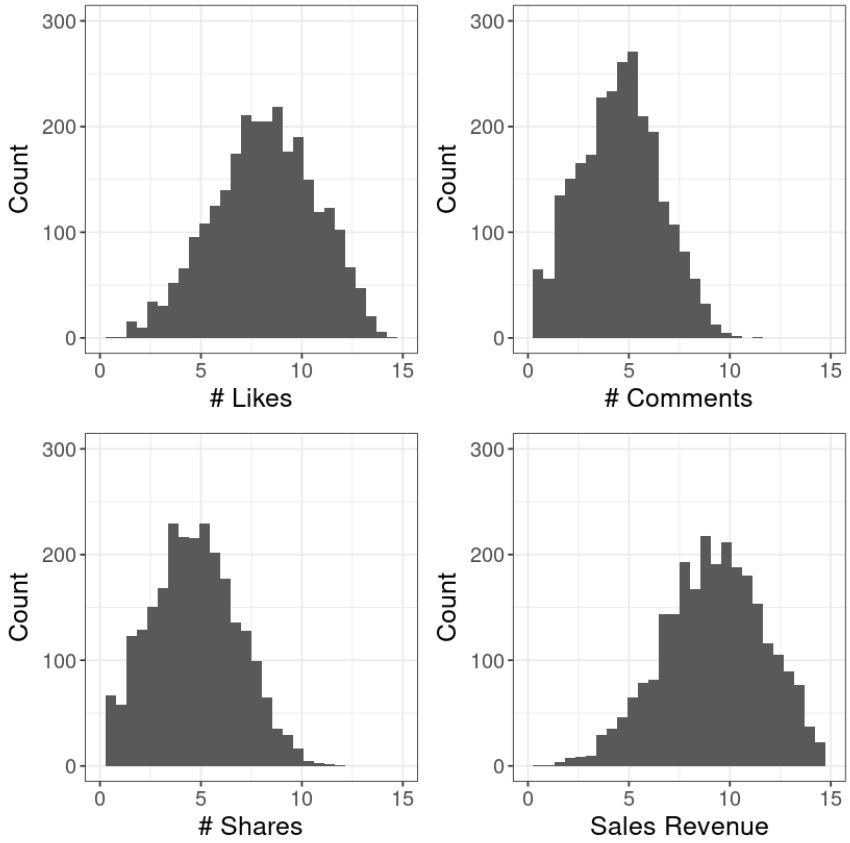
One challenge we face in our cross-category analysis is that 68% of the products in our sales panel miss category labels. Our solution is to predict missing category labels based on product titles. To do so, we draw on a sample of 8,447 products with category labels (including products outside the sales 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,[9] stopwords,[10] and chinese.misc[11] 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),[12] 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[13] in Caret.[14] 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.

[9] https://cran.r-project.org/web/packages/quanteda/quanteda.pdf.
[10] https://cran.r-project.org/web/packages/stopwords/stopwords.pdf.
[11] https://cran.r-project.org/web/packages/chinese.misc/chinese.misc.pdf.
[12] http://lsa.colorado.edu/papers/dp1.LSAintro.pdf.
[13] https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf.
[14] https://cran.r-project.org/web/packages/caret/caret.pdf.

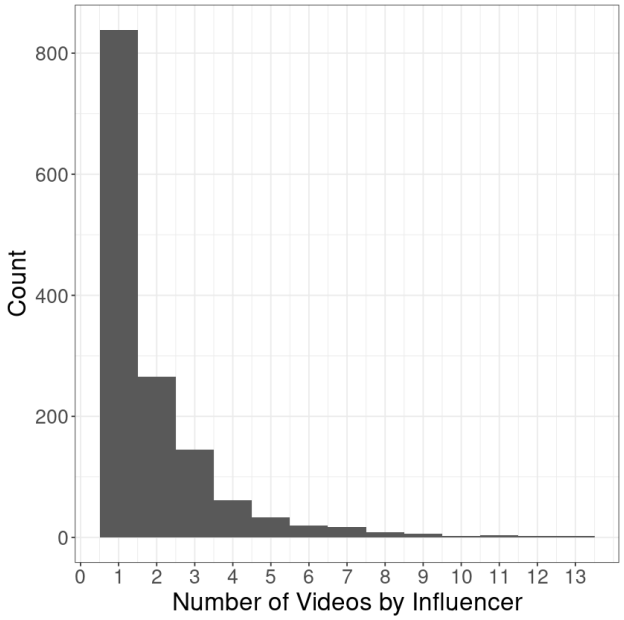
F Additional Summary Statistics of the Sales Panel

Figure F.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 30-day sales revenue (product level), all in log scale.

Figure F.2. Distribution of the Number of Videos by Influencer



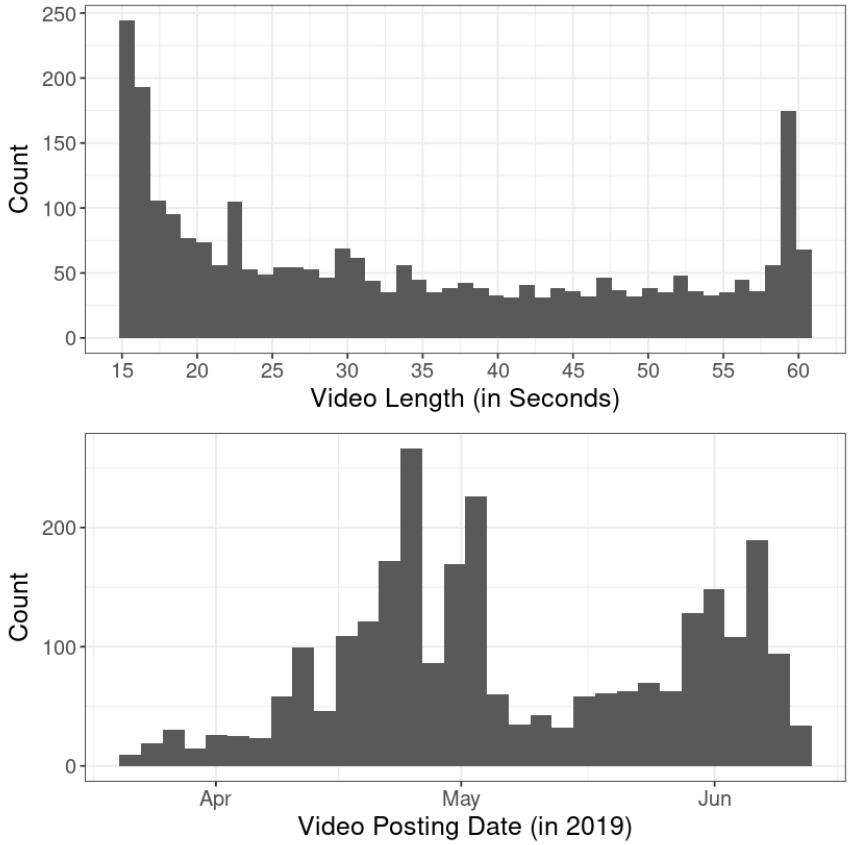
Note: The sample consists of all videos in the sales panel. Each observation is a video.

Table F.1. Summary Statistics of Influencer Characteristics

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

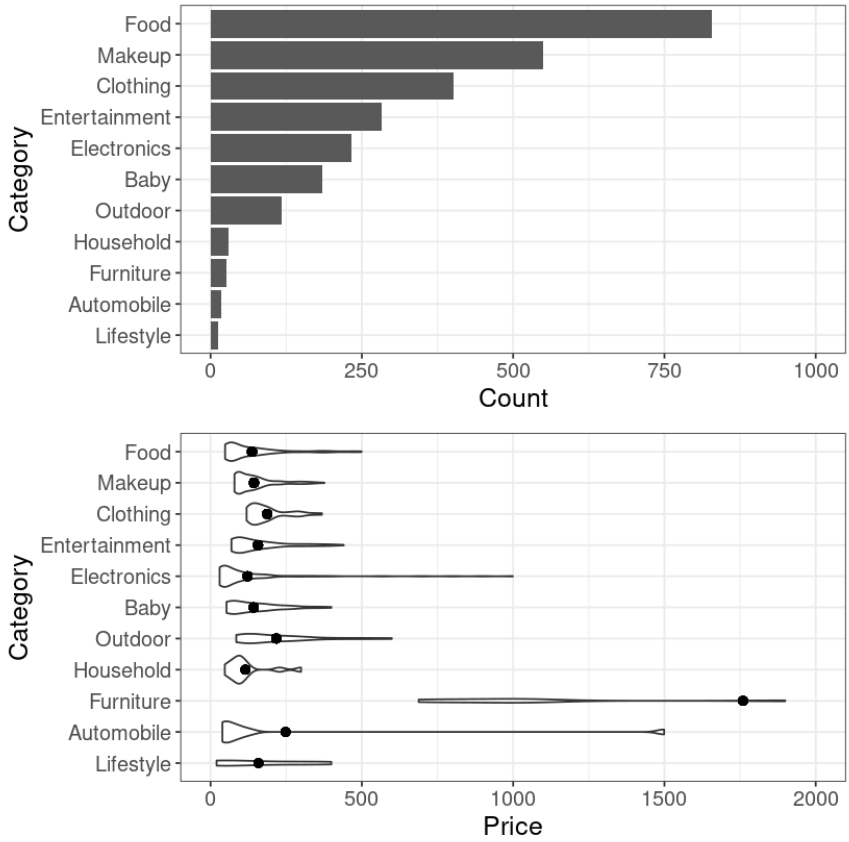
Note: The sample consists of all influencers in the sales panel. Each variable is at the influencer level. All observations were recorded in January 2019. For gender, 0 denotes female and 1 denotes male. CPM is the cost per 1,000 plays. Price per video ad is in RMB.

Figure F.3. Distribution of Video Length and Posting Time



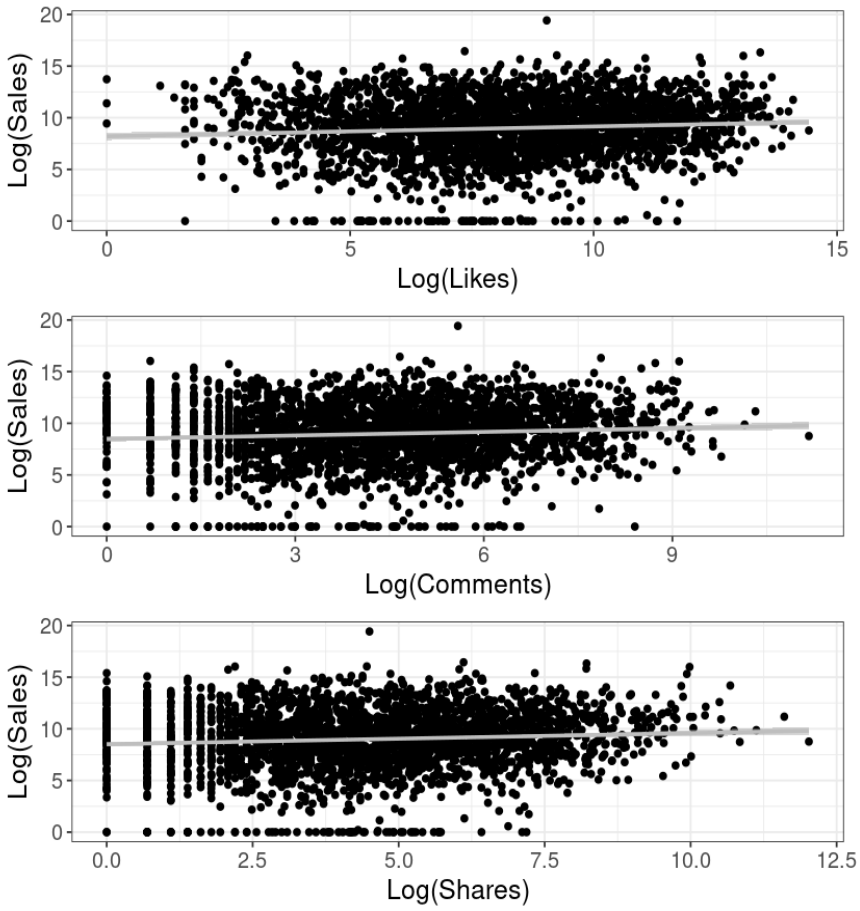
Note: The sample consists of all videos in the sales panel. Each observation is a video.

Figure F.4. Distribution of Product Category and Price Range



Note: The bottom panel is a violin plot of price distribution by category. Dots represent mean prices in category. Price is in RMB. The sample in the bottom panel consists of products in the sales panel with the highest and lowest 5% by price in each category removed for visualization.

Figure F.5. Correlation between Engagement and Sales



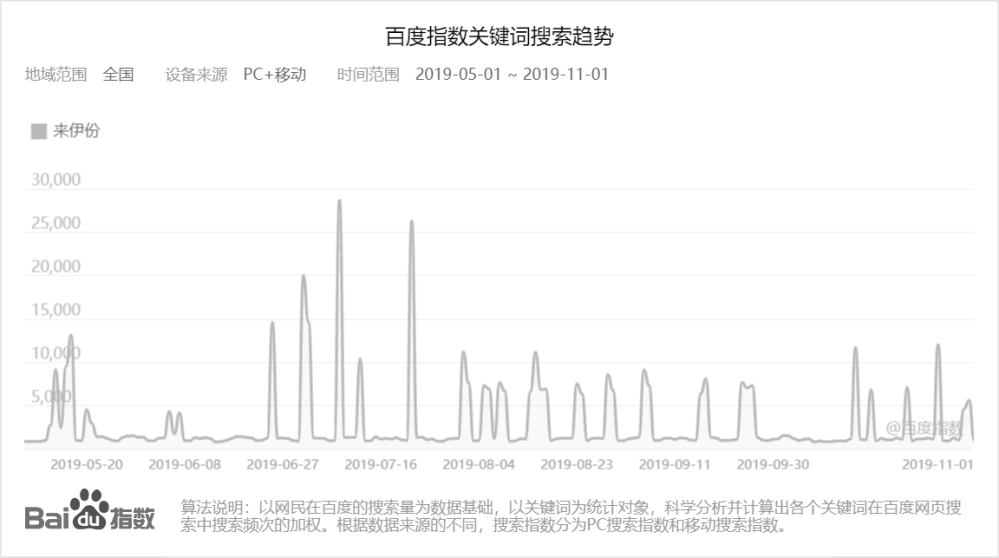
Note: The sample consists of all videos/products in the sales panel. The scatter plots present the relationship between raw video-level engagement measures (the number of likes, comments, and shares) and product sales revenue, both in the logarithmic scale to control for outliers. Sales has no significant correlation with any of the raw engagement metrics (ρ = 0.0022, p = 0.91 for likes; ρ = 0.0024, p = 0.90 for comments; ρ = 0.00093, p = 0.96 for shares).

G Baidu Search Index

As a proxy of unobserved time-varying demand, we collected data on the Baidu search index for all 2,685 products in our sales panel. Two research assistants manually entered the brand of each product as the keyword to track on the Baidu Index website (batch data collection is not available). Baidu Index currently accommodates keyword search at the level of product brand, not specific products. Nevertheless, we expect the search results to capture unobserved demand shifters such as brand campaigns or product campaigns that generate spillover effects within the same brand.

The scope of keyword search was set to include queries from all over the country (China), from both personal computers and mobile devices, and from May 1 through November 1, 2019 to match the time frame of the sales panel. Figure G.1 presents an example of search results from one keyword.

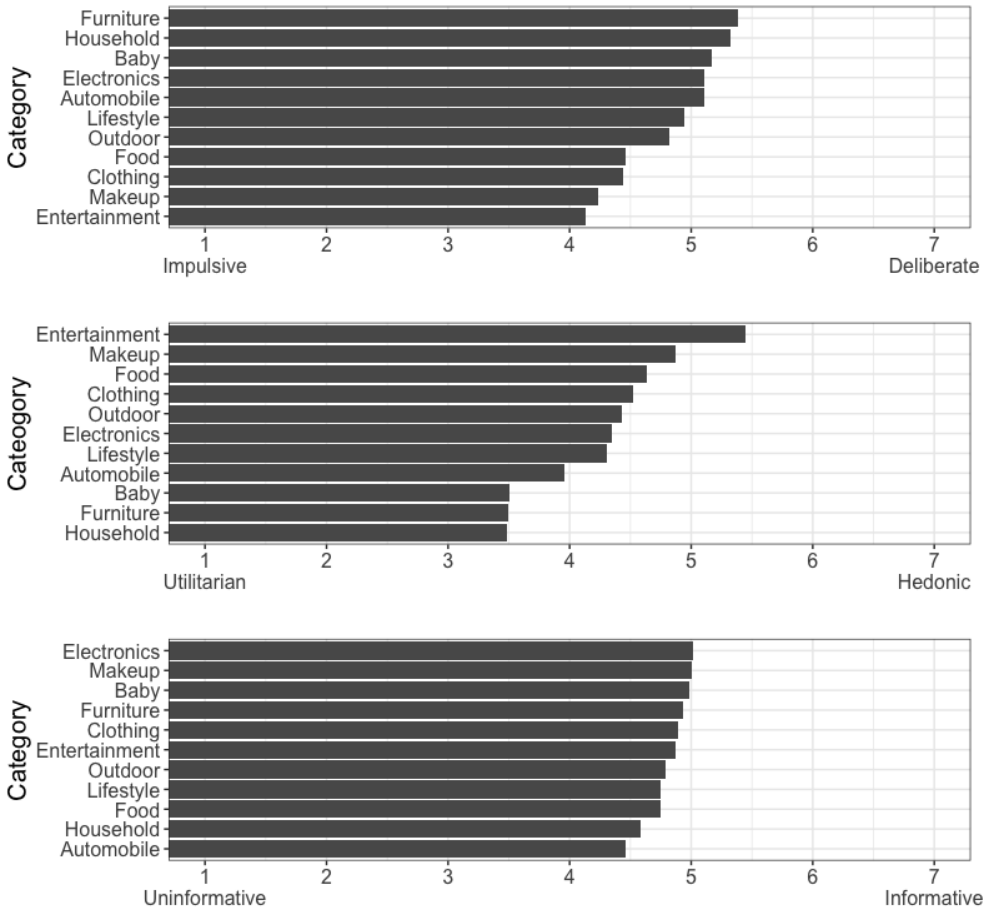
Figure G.1. Baidu Search Index Screenshot



Out of all products in the sales panel, we were able to obtain Baidu search results of 429 products. Visual inspection suggests that these tend to be products from bigger, more recognizable brands. For products without search results, we treat their search data as sequences of zeros. Replacing zeros with other constants does not affect our test of interest because we include product fixed effects in the analysis.

H User Survey of Product Category Perception

Figure H.1. Mean Ratings by Product Category



I Robustness Checks and Extensions

In this section, we extend the analysis presented in the main text to check the robustness of our algorithm. We check broadly two aspects of robustness, with respect to the construction of the algorithm, and with respect to the causal identification of sales lift.

I.1 Alternative Construction of the Engagement Heatmap

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 retrain the algorithm using the number of likes and comments instead. We re-estimate the main specification based on these alternative measures of raw engagement. Columns (1) and (2) of Table I.1 present the results. The moderation effect of pe-score on sales lift remains positive and significant when raw engagement is measured by the number of likes, but becomes insignificant when raw engagement is measured by the number of comments. One possible reason is that shares and likes are stronger indicators of positive engagement while comments can have different valences (e.g., Azer and Alexander 2020).

So far, we have followed the "supervised" approach to construct the 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 supervised 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 engagement using the "unsupervised" approach.

The motivation of the unsupervised approach is that 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

Table I.1. Alternative Measurement of Engagement Heatmap

	Sales Revenue		
	Likes	Comments	Unsupervised
Post	−98.55	−49.92	−64.36
	(82.25)	(84.20)	(83.05)
Post × PE-Score (# Likes)	789.67***		
	(160.82)		
Post × PE-Score (# Comments)		172.36	
		(172.60)	
Post × PE-Score (Unsupervised)			323.86*
			(160.00)
Post × Engagement	−46.04	−39.95	−31.55
	(164.14)	(165.14)	(164.65)
Post × Product	−721.01	−471.19	−588.49
	(388.27)	(385.33)	(391.00)
Post × Engagement × Product	353.58	463.68	494.97
	(815.57)	(818.63)	(815.05)
Search	7.43**	7.38**	7.41**
	(2.76)	(2.76)	(2.76)
Post × Covariates	Yes	Yes	Yes
Video/Product, Influencer, Day FE	Yes	Yes	Yes
Observations	173,515	173,515	173,515
Adjusted R^2	0.98	0.98	0.98

Note: See Table 4 note.

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.[15] As column (3) of Table I.1 shows, the pe-score metric based on unsupervised learning continues to predict higher sales lift although the effect size and significance level is not as strong, which suggests that there is substantial value in collecting engagement data to construct a supervised algorithm.

[15]See Figure B.1b of the Online Appendix for an illustration and Hou and Zhang (2007) for more details of the algorithm.

I.2 Validity Check of the Engagement Heatmap

As discussed in the paper, the literature 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.[16] For each frame, the algorithm outputs the location of boxes that contain a human face. Similar to the product heatmap, we estimate a face 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 engagement values with the indicator variable of whether a face is present in a pixel. The result indeed shows a positive and significant correlation ($\rho = 0.04, p < 0.001$). This adds face validity (pun intended) to our engagement heatmap because we are now more confident that it is uncovering the more engaging parts of a video as we intended.

[16] https://pypi.org/project/face-recognition.

I.3 Pre-Trends, Dynamic Treatment Effects, and Dynamic Moderation Effects

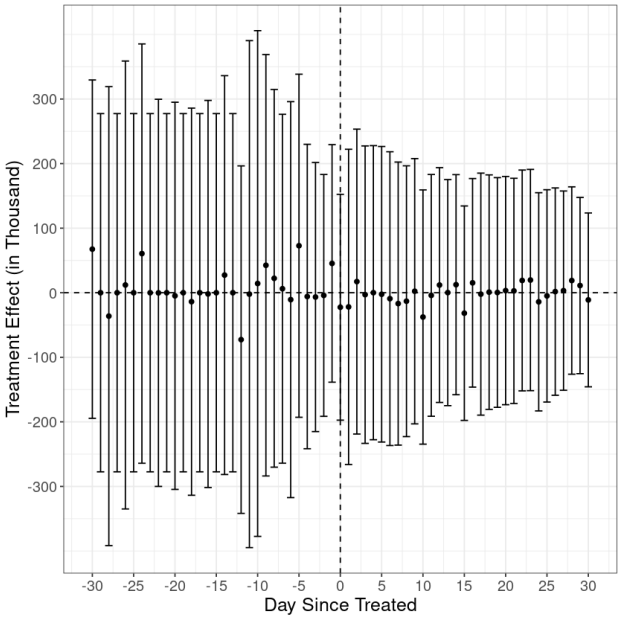
We first test for parallel pre-trends between treated and not-yet-treated products, an assumption that underlies the staggered diff-in-diff approach. We also investigate dynamic treatment effects to allow advertising effects to vary over time. To do so, for each product we include 30-day leads and lags 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:

$$\textit{Sales}_{vd} = \sum_{e=-30}^{30} \alpha_e \cdot \mathbf{1}\{d - t_{vi} = e\} + \textit{Video}_v + \textit{Influencer}_{i(v)} + \textit{Day}_d + \textit{Search}_{vd} + \epsilon_{vd}. \quad \text{(OA1)}$$

Figure I.1 presents the estimated coefficients α_e in 1,000 RMB 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 finding of an insignificant treatment effect from positing video ads, even if we allow the advertising effect to vary over time. This result also mitigates the serial correlation concern of diff-in-diff analysis (Bertrand et al. 2004).

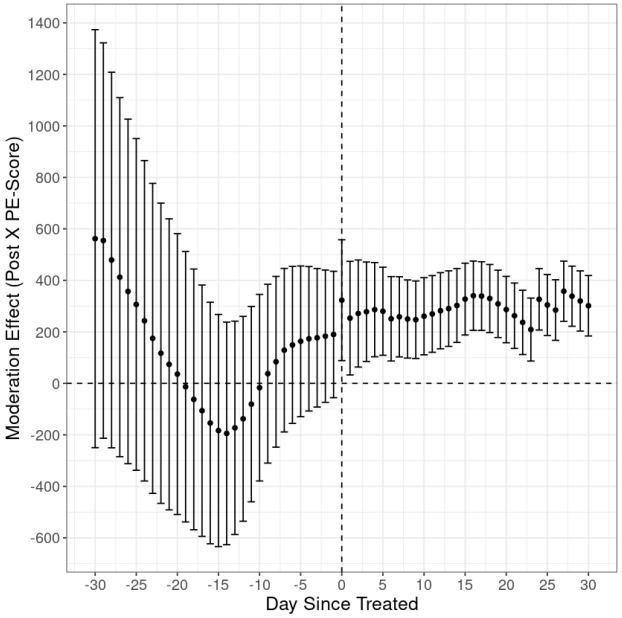
We then estimate a dynamic specification that adds the interaction between pe-score and the leads and lags of ad posting. Figure I.2 presents the dynamic moderation effects of pe-score in 1,000 RMB. Pe-score shows no significant moderation effects on sales prior to the post date, and positive and significant moderation effects that are persistent over 30 days after the post date. Given that our outcome variable is prior-30-day sales revenue, the dynamic effect e days ($1 \leq e \leq 30$) after ad posting should be interpreted as the cumulative effect on sales after e days rather than the effect on the e^{th} day.

Figure I.1. Pre-Trends and Dynamic Treatment Effects



Note: Bars are the 95% confidence intervals.

Figure I.2. Dynamic Moderation Effects



Note: Bars are the 95% confidence intervals.

I.4 Continuous Treatment

In our analysis thus far, treatment is coded as a binary variable of whether a video ad has been posted. We also test for a specification where treatment is coded continuously as cumulative ad exposure. The consideration of cumulative effects is particularly relevant in our setting because the dependent variable (DV) in the sales panel is a product's prior-30-day total sales revenue. For an ad posted within this 30-day window, the effect of ad exposure on the DV should strictly speaking be prorated based on the number of days the ad is present. For an ad posted before this 30-day window, the DV captures ad exposure over all 30 days. Finally, to capture the potential decay of ad effectiveness over time, we scale the effect of ad exposure on each day's incremental sales by a carryover factor, δ (Little 1979, Shapiro et al. 2021).

Specifically, we again use t_{vi} to denote the date when influencer i posts a video ad v. We define the product's cumulative ad exposure as of day d, A_{vd}, as:

$$A_{vd} = \begin{cases} 0 & \text{if } d - t_{vi} < 0 \\ \sum_{l=0}^{d-t_{vi}} \delta^l & \text{if } d - t_{vi} \in [0, 30) \\ \sum_{l=0}^{29} \delta^{l+(d-t_{vi}-29)} & \text{if } d - t_{vi} \geq 30. \end{cases} \tag{OA2}$$

We re-estimate Equations 2 and 3 with the binary treatment $Post_{vd}$ replaced by A_{vd}. Following standard practice (e.g., Shapiro et al. 2021), we repeat the estimation varying δ from 0 to 1 at increments of 0.1, where $\delta = 0$ represents the case of no ad effect carryover and $\delta = 1$ the case of no ad effect decay. As shown in Table I.2, the main effect of ad exposure is insignificant whereas the moderation effect of pe-score remains positive and significant for all values of δ except $\delta = 1$. Our results are thus robust unless ad effect carries over perfectly, which may not be realistic for many products on TikTok (see also Figures I.1 and I.2). Note that we could not estimate δ with our data via goodness-of-fit because the residual variation after controlling for fixed effects is too small.

Table I.2. Continuous Treatment and Moderation Effects

Carryover Factor (δ)	Ad Exposure	Ad Exposure $\times$ PE-Score
0.0	-4.37	91.87^{**}
	(5.80)	(28.56)
0.1	-3.90	83.63^{**}
	(5.25)	(25.76)
0.2	-3.41	75.30^{**}
	(4.70)	(22.97)
0.3	-2.91	66.85^{***}
	(4.15)	(20.17)
0.4	-2.44	58.25^{***}
	(3.59)	(17.37)
0.5	-1.98	49.42^{***}
	(3.03)	(14.57)
0.6	-1.55	40.24^{***}
	(2.46)	(11.77)
0.7	-1.11	30.52^{***}
	(1.89)	(8.98)
0.8	-0.64	20.01^{**}
	(1.32)	(6.23)
0.9	-0.31	8.34^{*}
	(0.76)	(3.65)
1.0	1.10^{**}	-3.88^{*}
	(0.37)	(1.72)
Observations	173,515	173,515
Adjusted R^2	0.98	0.98

Note: The sample is the sales panel. Each observation is at the product-day level. OLS. Dependent variable is a product's prior-30-day sales revenue in 1,000 RMB. Main effects are estimated from Equation 2 and moderation effects from Equation 3 with the binary treatment Post replaced by the continuous treatment Ad Exposure as given in Equation OA2. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

I.5 Simulation of Different Treatment Specifications

To better understand the impact of different treatment specifications, we conduct a simulation study that mimics the setup of our sales panel. We simulate a panel with $N = 1,000$ products over $D = 100$ days. We simulate one ad for each product. Ad posting time is random and uniformly drawn within the 100 days. The true treatment

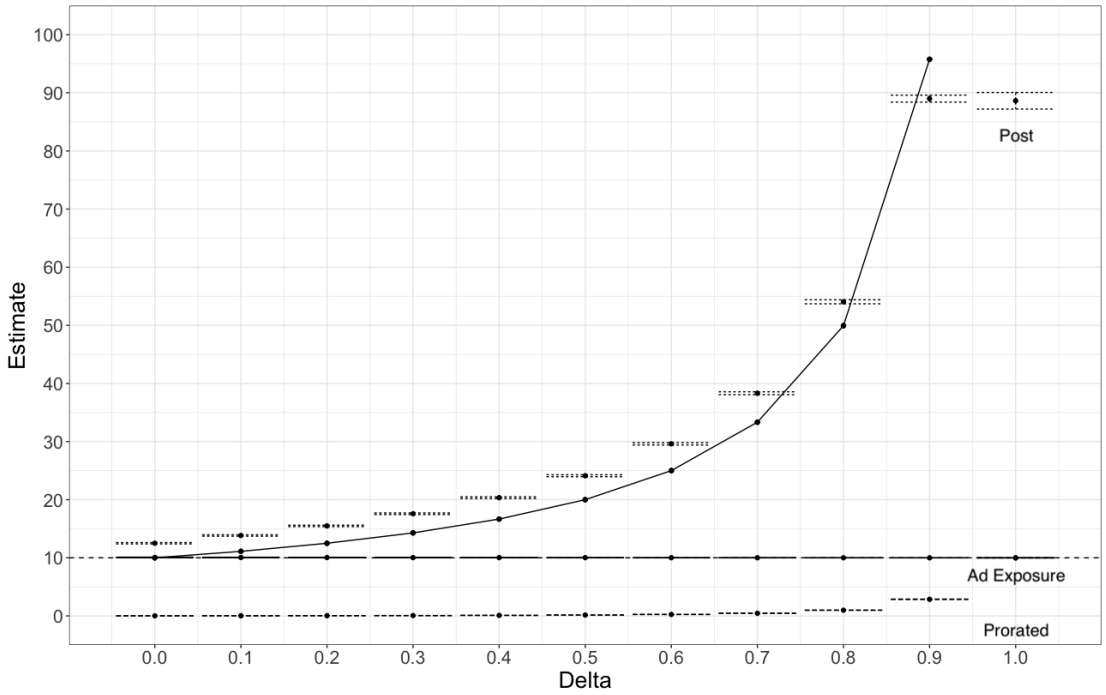
effect of ad v on sales revenue of day d is:

$$\tau_{vd} = \begin{cases} 0 & \text{if } d - t_v < 0 \\ \tau \cdot \delta^{d-t_v} & \text{if } d - t_v \geq 0. \end{cases} \tag{OA3}$$

The parameter τ is the true treatment effect on the day an ad is posted, which we set to be 10 in the simulation. Prior-30-day sales revenue in the absence of treatment is simulated as $Y_{vd} = v + \sum_{l=\max(0,d-29)}^{d} l + \epsilon_{vl}$ to capture a product-specific effect, a simple linear time trend, and a random normal error term that is correlated within a product over time. We set the correlation coefficient to be 0.5; simulation results are robust with a correlation coefficient of 0.1. We regress prior-30-day total sales revenue on treatment controlling for product and day fixed effects. Specifically, we consider treatment coded as (1) the binary variable Post as in the paper, (2) the continuous variable Ad Exposure as defined in Equation OA2, and (3) a nested version of Ad Exposure where $\delta = 1$, which measures the number of days an ad is posted in the last 30 days (called "Prorated").

Figure I.3 shows the simulation result. We plot the true 30-day total treatment effect for δ from 0 to 0.9 (the effect when $\delta = 1$ is 300 and is removed from the plot to facilitate visualization). Post closely matches the trajectory of the true total 30-day treatment effects up to $\delta = 0.9$. Ad Exposure correctly identifies the treatment effect for all values of δ. Prorated underestimates the treatment effect except when there is no ad decay, in which case it is equivalent to Ad Exposure by definition.

Figure I.3. Simulation Result of Different Ad Treatment Specifications



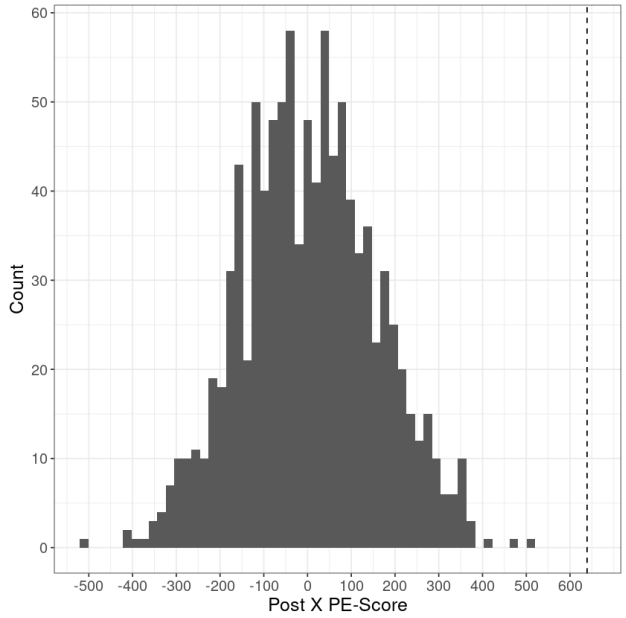
Note: The y-axis is the estimated advertising treatment effect. Delta (δ) is the ad effect carry-over factor. The true treatment effect on the day an ad is posted is set to be 10 in the simulation. The curve is the true 30-day total treatment effect for δ from 0 to 0.9. The true 30-day total treatment effect when $\delta = 1$ is 300 and is removed from the plot to facilitate visualization. Bars are the 95% confidence intervals.

I.6 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 pe-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 pe-score. As shown in Figure I.4, these estimated effects are below the effect from the main analysis in the paper (the dashed vertical line) and the difference is significant ($p < 0.001$); they are also not significantly different from zero ($p = 0.36$). This result gives us further confidence that the moderation effect of pe-score we have uncovered is not due to chance and indeed captures the predictive power of our algorithm.

Figure I.4. Randomization Inference for the Moderation Effect of PE-Score



Note: The figure presents the distribution of the coefficients of Post × PE-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.

I.7 Alternative Diff-in-Diff Specifications

A recent stream of methodological papers on diff-in-diff argue that, when the adoption of the treatment by different groups is staggered over time and treatment effects vary across groups or over time, the standard two-way fixed effects (TWFE) specification might not identify an aggregate effect that can be readily interpreted as the average treatment effect on the treated (e.g., Goodman-Bacon 2018, De Chaisemartin and d'Haultfoeuille 2020, Sun and Abraham 2020, Callaway and Sant'Anna 2021).

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"). Though negative weights are not problematic when treatment effects are homogeneous, the TWFE estimator can be severely biased when treatment effects are heterogeneous – particularly when they change over time within treated units. This can cause issues such as the treatment effect being positive for all units with TWFE giving a negative result.

We first test for how pervasive the negative weights are in our setting. By applying the Frisch-Waugh-Lovell theorem, the TWFE estimator can be summarized as:

$$\beta_{\text{TWFE}} = \sum_{vd} Sales_{vd} \cdot \frac{\tilde{Post}_{vd}}{\sum_{vd} \tilde{Post}_{vd}^{2}}$$

where $\tilde{Post}_{vd}$ is the residual from a regression of the treatment indicator on the video/product, influencer, and day fixed effects (De Chaisemartin and d'Haultfoeuille 2020). We compute these weights for all treated product-day combinations and the negative weights account for about 12% of the observations which is a relatively small fraction (Jakiela 2021).

While the literature has not settled on a standard approach, we test the robustness of our result with two TWFE alternatives – two-stage diff-in-diff (Gardner 2022) and diff-in-diff implemented via an explicit imputation procedure (Borusyak et al. 2022).

Gardner (2022) develops a two-stage regression approach to identification that is ro-

bust to treatment effect heterogeneity when treatment adoption is staggered. The first stage of the procedure consists of a regression of sales on a set of controls, which in our setting include video/product, influencer, and day fixed effects, as well as the search index. In the second stage, the estimated effects from the controls are subtracted from observed sales, and these residualized sales are regressed on the treatment status. Under the usual parallel trends assumption, this procedure identifies the overall average treatment effect on the treated across products and time periods, even when treatment effects are heterogeneous over groups and periods.

Borusyak et al. (2022) proposes an intuitive imputation-based procedure that does not restrict treatment effect heterogeneity. Counterfactual sales in the absence of the treatment is imputed based on the controls, which again include video/product, influencer, and day fixed effects as well as the search index. After deriving counterfactual sales for each treated product-day combination, the product-day level treatment effect is simply estimated as the difference between observed and imputed sales. An average treatment effect on the treated is then computed by aggregating these product-day level treatment effects.

We follow these two approaches and estimate an average treatment effect on the treated separately for the full sample and the subsample of products with pe-score in the top 10% range. The computation is performed using the did2s package[17] for Gardner (2022) and didimputation[18] for Borusyak et al. (2022) in R.

As Table I.3 shows, the results are consistent with those in the paper: the treatment effect of posting a video ad is insignificant on the full sample of videos but significantly positive for videos with high pe-scores.

[17] https://kylebutts.com/did2s.
[18] https://github.com/kylebutts/didimputation.

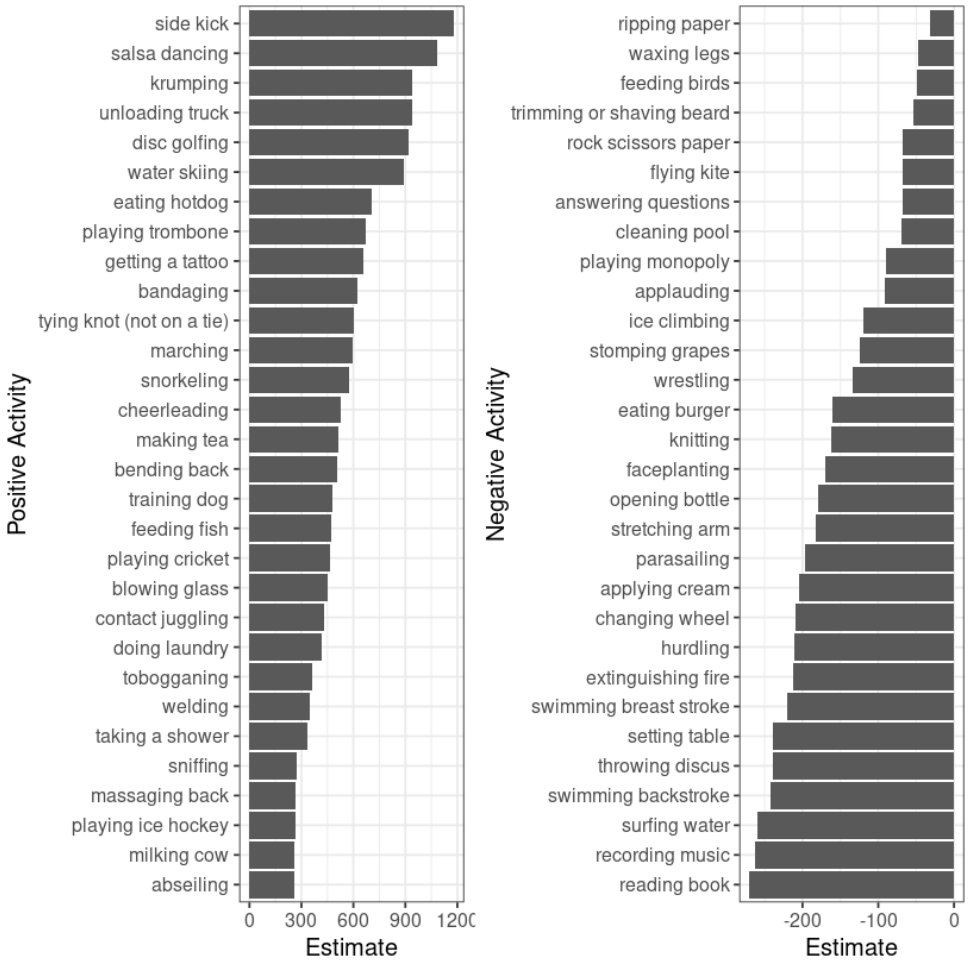
Table I.3. Alternative Diff-in-Diff Specifications

	Sales Revenue			
	Gardner (2022)		Borusyak et al. (2022)	
	All Video Ads	High-PE-Score Ads	All Video Ads	High-PE-Score Ads
Post	-4.60	416.01***	23.74	606.91*
	(8.50)	(67.09)	(59.46)	(281.63)
Controls	Yes	Yes	Yes	Yes
Observations	173,515	17,072	173,515	17,072

Note: Full sample is the sales panel. High pe-score means a subsample of videos/products in the top 10% of the pe-score distribution. Gardner (2022) and Borusyak et al. (2022) specifications. Sales revenue is in 1,000 RMB. Controls include video/product, influencer, and day fixed effects and the search index. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

J Activities and Engagement

Figure J.1. Activities and Engagement



Note: Results are relative to a baseline where no activity is detected.

Table J.1. Activities and Engagement: Top 10 Words in the Topic Model

Positive Activities		Negative Activities	
Topic 1	Topic 2	Topic 1	Topic 2
play	make	swim	paper
clean	dance	wax	rip
ice	eat	scissor	leg
ski	back	fli	bird
dog	fish	kite	feed
car	blow	answer	beard
hair	push	question	shave
fold	climb	clean	trim
head	floor	split	rock
kick	basketbal	play	pool