Engagement that Sells: Influencer Video Advertising on TikTok

Submitted to *Marketing Science*

July 11, 2023

Abstract

Many ads are engaging, but what makes them engaging may have little to do with the product. We call this ineffective engagement. This problem is exacerbated in influencer advertising when the incentives of influencers and advertisers are not perfectly aligned. We develop an algorithm to measure the effective engagement and predict the sales lift of influencer video advertising. We propose the concept of product engagement score, or PE-score, to capture how engaging the product is as presented in a video. We estimate pixel-level engagement as a saliency map by training a deep 3D convolutional neural network on video-level engagement data and locate pixel-level product placement with an object detection algorithm. PE-score is computed as the pixel-level, engagement-weighted product placement in a video. We construct and validate the algorithm with influencer video ads on TikTok and product sales data on Taobao. We leverage variation in video posting time to identify video-specific sales lift and show that the PE-score strongly and robustly predicts sales lift. We discuss how various stakeholders in influencer advertising can use the PE-score in a scalable way to evaluate content, align incentives, and improve efficiency.

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

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-form 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 has reached 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 become the biggest video-based influencer advertising channel by 2024 and the second biggest overall only behind Instagram.[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, comments, and shares) 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 lift 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. We call this ineffective engagement. 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 for effective engagement, a metric that captures the extent to which engagement is driven by the product, or the ad is engaging for the "right" reason. We call this metric *product engagement score*, or *PE-score* for brevity.

We operationalize the 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 space and time of a video in which a product is presented. We compute the PE-score in three steps.

First, we construct a three-dimensional (3D) *engagement heatmap* for each video to measure the importance of each pixel to overall video-level 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 (CNN) using video-level engagement data. A video's engagement heatmap is then derived as a pixel-level

[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 a disconnect between brands and influencers, *OST*, January 2, 2019.

saliency map, which outputs the gradient of video-level engagement with respect to each pixel in the video.

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 (SIFT)."

Third, we compute the 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 equivalently, 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 dataset of 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 sales lift. This observation 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, the PE-score tends to be higher in these videos than in videos where influencers advertise for another party.

[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.

For our main test, we first estimate video-specific sales lift of influencer video ads via the difference-in-differences (DID) method, leveraging the variation in video posting time for causal identification. We then explore predictors of video-specific sales lift. Consistent with our hypothesis, the PE-score is a strong and robust predictor of sales lift. Notably, overall video engagement or product placement does not predict sales lift. Moreover, being both engaging and intensive with product placement but doing so separately does not predict sales lift either. These results highlight the unique predictive power of the PE-score – simply making the video more engaging or featuring the product more may not help; it is product engagement that drives sales.

For more actionable insight, we further explore possible drivers of the 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. Aligning product placement spatiotemporally with these elements of engagement might help enhance sales conversion.

The PE-score can be practically valuable in several ways. First, we invested heavily in algorithm calibration, such that the 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 to release. Second, the 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 the 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.

[9]The product image can be sourced from an e-commerce website where the product is being sold but can also simply be a cropped image from the video when the product is shown.

2 Related Research

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 29% 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.

Research on influencer marketing has studied a range of topics including influencer dissemination of ads (Gong et al. 2017), consumer trust (Lou and Yuan 2019), influencer versus celebrity endorsements (Schouten et al. 2020), influencer selection (Valsesia et al. 2020, Tian et al. 2022), cultural effect (Bentley et al. 2021), drivers of engagement (Leung et al. 2022), returns on influencer promotions (Huang and Morozov 2022), and the impact of influencer posts on copyrighted content (Li et al. 2023).

We contribute by studying one of the latest forms of influencer marketing – influencer video advertising. With the rapid growth of video as a communication tool, influencer video advertising is gaining popularity in practice and attention in academia. One recent paper particularly related to ours is Rajaram and Manchanda (2023), which developed an interpretable deep learning framework to study the impact of various video-ad modalities (i.e., text, audio, image) on engagement. We also analyze influencer video ad content but focus on sales conversion – we develop an algorithm to predict sales lift from ad content.

Our focus on sales lift adds to the discussion of ad engagement versus conversion. A series of papers have found that engagement does not guarantee conversion. Ad viewing may even be negatively related to buyer intent (Teixeira et al. 2014, Tucker 2015), virality does not always add value to the brand (Akpinar and Berger 2017), Facebook liking has

[10] The state of influencer marketing 2023: benchmark report, *Influencer Marketing Hub*, February 7, 2023.

no positive impact on consumer attitudes or purchases (John et al. 2017), and engaging contents (consumer selfies) may not improve purchase intentions (Hartmann et al. 2021). 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] An established stream of research links video content with real-time viewer behaviors during the process of video consumption. Various measurement innovations have been developed, including handheld devices (Polsfuss and Hess 1991), "feeling monitor" computers (Baumgartner et al. 1997), eye tracking (Wedel and Pieters 2008, Teixeira et al. 2010), electroencephalography (Barnett and Cerf 2017), facial expression tracking (Liu et al. 2018), functional Magnetic Resonance Imaging (Tong et al. 2020), and viewer live comments analysis (Zhang et al. 2020).

We contribute to this video-content-design literature along four dimensions. First, the literature has focused on movies or standard video ads. We study a new type of content – video ads produced by influencers. Influencer video ads can be fundamentally different from traditional video ads. In particular, influencers' incentives to promote themselves may affect ad content design. Second, many methods proposed in this literature rely on real-time viewer behaviors data for new videos 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 can be

[11] Another closely related, growing 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. 2022), logo design (Dew et al. 2022), labor-market research (Troncoso and Luo 2022), product-returns management (Dzyabura et al. 2023), and business-survival prediction (Zhang and Luo 2023).

applied directly and in a scalable way 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 more-granular, spatiotemporal approach to video content design helps reveal further 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 for 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 the PE-score is that, other things being equal, the more engaging an advertised product is in an influencer video ad, the more 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.

[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.

[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.

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 the 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 matrix 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 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 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 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 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 en-

[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 a top-5 accuracy of 0.95.

gagement. Shares can be a stronger signal of engagement than likes and comments, as users are willing to endorse shared videos on their social networks.[16] Shares are also a common subject of academic research (e.g., Akpinar and Berger 2017, Tellis et al. 2019). However, our results are robust if we use likes or comments to measure engagement (see Online Appendix H.1).

Prior to training, we regress video-level raw engagement data on influencer fixed effects, product fixed effects, acoustic features, and transcript embeddings.[17] We retain the residuals from the regression and use them as labels to train the 3D CNN. Using engagement residuals instead of raw engagement allows us to not only control for outliers but also focus on the variation in engagement that is driven by the visual component of the video ad holding other features that might affect engagement fixed.[18] 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, all randomly chosen.

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

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

[17] The non-visual features can be used as side features and combined with visual features for joint training. This would account for non-linearity and the interaction between visual and non-visual features. We follow the residual approach to train the 3D CNN only on visual features for the ease of extracting saliency maps. See Fong et al. (2021) for an in-depth study of music and Rajaram and Manchanda (2023) for joint analysis of multiple video elements.

[18] For example, it is important to note that the 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.

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.[19] 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.[20] We train the 3D CNN on a high-performance computing (HPC) cluster using TensorFlow.[21] 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 a 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, 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. We

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

[20] There is a paper famously titled "I just ran four million regressions" (Sala-i-Martin 1997). Here we just ran one regression with over seven million parameters.

[21] https://keras.io.

ignore the sign per the standard implementation of saliency maps because derivatives are with respect to increasing the value of a pixel along a particular color channel, which means increasing its intensity or brightness and does not have an intrinsic, meaningful interpretation.[22] 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 image 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.[23] See Online Appendix B for more details on the saliency map.

To summarize the engagement heatmap from an 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 across space and time.[24] 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 spoken content.

[22] For an interactive example, see http://www.cknuckles.com/rgbsliders.html.

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

[24] Similar to standard regressions, this approach requires sufficient spatiotemporal variation in pixel values and relies on functional-form assumptions. We are likely to have sufficient pixel variation given the high variety of videos on TikTok. 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 matrix 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.[25]

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.[26]

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 matrix of binary values, where 1 indicates product presence at a pixel and 0 indicates absence. See Online Appendix C for more details on the product heatmap.

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

[26] 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 a cutoff value of 0.75 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.

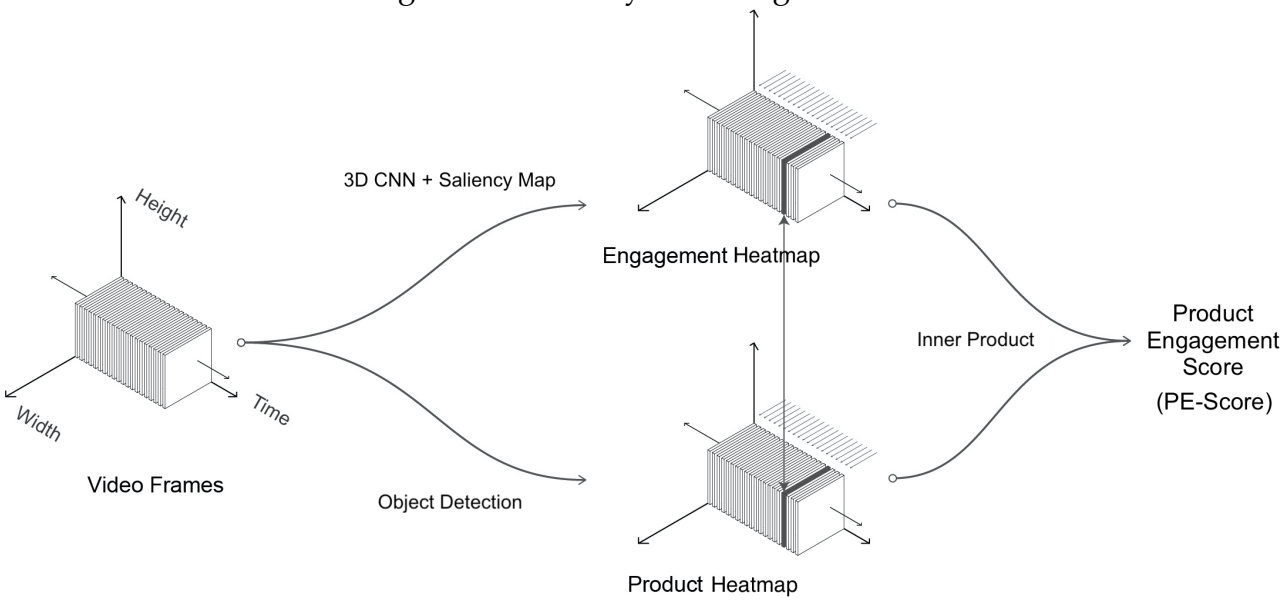
3.3 Computing PE-Score

In the third step, we combine the engagement heatmap and the product heatmap to calculate the 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 the total number of pixels, or the volume, of video v. As discussed, we interpret the 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.

Figure 1. Summary of the Algorithm



Two remarks on our algorithm are in order. First, we train a 3D CNN on video content data using video-level engagement as the outcome variable. The number of parameters to estimate far exceeds the training sample size. This is a common feature of deep learning models and does 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, as we discuss in Online Appendix A, the 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 the 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, the PE-score is constructed without sales data and is evaluated on its ability to predict sales out of sample. This is arguably a more stringent test than testing the predictive power of a model that uses sales as the outcome variable both in and out of sample. Third, we do not observe the sales lift of each video ad in the data; it needs to be estimated via a DID approach, as we will explain. The number of videos with sales data we have (see next section) may not be sufficient to train a complex 3D CNN. We leave the training of end-to-end sales models as a topic for future research.

In what follows, we evaluate our algorithm with sales data. Specifically, we test whether influencer video ads with 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 revenue 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.[27]

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 2 million influencers and 1.9 million registered sellers.[28] 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

[27] For instance, Lee et al. 2018 studied advertising content and engagement on Facebook, and highlighted the lack of access to sales data as a limitation.

[28] The way up for 2 million Xingtu talents, *Trend Insight*, August 25, 2022.

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.[29] 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, as well as influencer characteristics.

The second source is a product dataset that contains all products on Taobao that were advertised on TikTok 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 stated as "30-day sales," meaning the sum of sales revenue over the previous 30 days, including the current day, measured at the daily level.[30] 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.[31]

[29] Top 15 Chinese E-commerce websites in 2023, *TMO Group*, February 17, 2023.

[30] There are some missing sales observations for technical reasons that our partner company believed to have occurred randomly. As a result, we have an imbalanced sales panel.

[31] 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.[32] Among influencer video ads from the first source, 2,734 have matching product sales data. Among these 2,734 video ads, 2,685 have complete influencer characteristics. We call the panel dataset of these 2,685 product-video pairs the *sales panel*, which we will set aside to evaluate 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 the 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 a holdout test of the algorithm, all through random assignment. We call the cross-sectional dataset of these 16,951 video ads the *construction sample*. Altogether, this paper draws on a total of 19,636 video ads.

Note that we do not require the sales panel and the construction sample to be comparable. In fact, once the algorithm is constructed, PE-scores of videos in the sales panel can be computed prior to release without relying on their engagement or sales data. This feature contributes to algorithm scalability. It also offers a stringent test of our algorithm – we construct it from one sample of videos without any sales information (the construction sample) and test its predictive power on sales lift on a different sample (the sales panel), which helps examine the external validity of the algorithm.

Table 1 presents summary statistics of observed video engagement metrics: the number of likes, comments, and shares.[33] Engagement takes time to grow. To capture each video's ultimate level of engagement, we use its last observed value in our data, which occurs, on average, 28 days after posting. These engagement metrics are statistically indistinguishable between the sales panel and the construction sample, except that the former were shared less on average.

[32] Some products might have changed their ID during the data window, which prevented us from matching every video to the corresponding product. More generally, the sales panel may not be a random subset of products on Taobao. We expect the effect magnitude of the PE-score to depend on the specific product market. However, we construct our algorithm based on all available TikTok influencer video ads that have spoken words during our data window. Therefore, our *algorithm* can plausibly generalize.

[33] We also observe the number of plays for each video. Play volume can be a noisy measure of engagement because it does not capture how much time users actually spend on a video. We control for play volume in subsequent analysis.

Table 1. Summary Statistics of Observed Video Engagement

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 in the Construction Sample (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. Videos in the sales panel do not overlap with videos in the construction sample.

The top of table 2 presents summary statistics of sales revenue, as well as prices and discounts, of all products in the sales panel. The average 30-day sales revenue of products in our data is 246,680 RMB, or 35,699 USD based on the average 2019 currency exchange rate of 6.91:1. The bottom of Table 2 presents summary statics of all influencers in the sales panel. Each influencer posted 1.9 video ads in the sales panel on average.

Table 2. Summary Statistics of Products and Influencers (Sales Panel)

Variable	N	Mean	St. Dev.	Min	Median	Max
Products						
Average 30-Day Sales	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
Influencers						
Gender (0=Female, 1=Male)	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	635,432	3,255,567	0	74,908	97,890,191
Price per Video Ad	1,404	19,530	53,808	0	6,000	1,000,000
Expected CPM	1,404	1,027	21,315	0	121	785,714
# Video Ads Influencer Has Posted	1,404	13	27	0	2	265

Note: Each product-related variable is at the product level and is measured in RMB. A product's average 30-day sales revenue is taken over its observed days in the sales panel. Price and discount contain no variation at the product level over the duration of our data. Each influencer-related variable is at the influencer level and was recorded in January 2019. CPM is the cost per mille (1,000) plays in RMB. Price per video ad is in RMB.

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). The most common video length is 15 seconds whereas the video posting date is widely distributed (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 notable pattern in the data is the lack of a significant correlation between video engagement and sales lift. Figure F.5 presents the scatter plots of the relationship between raw engagement metrics (the number of likes, comments, and shares) and the difference in average 30-day sales before and after a video ad is posted. Sales difference has no significant correlation with the raw engagement metrics ($\rho = -0.0075$, $p = 0.90$ for likes; $\rho = -0.02$, $p = 0.74$ for comments; $\rho = -0.0074$, $p = 0.91$ 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

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. Last, 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 video-level summary statistics of these three outputs for videos in the sales panel, which is the sample we will use to evaluate the algorithm. In the table, the "engagement score," termed to differentiate it from actual engagement, is a video's sum of pixel-level engagement values. The "product score" is a video's sum of pixels in which the product appears. To facilitate interpretation, we

normalize all three scores to the interval of $[0, 1]$ in this table and in subsequent analysis.

Table 3. Summary Statistics of the Video-Level Computed Engagement Score, Product Score, and PE-Score (Sales Panel)

Variable	N	Mean	St. Dev.	Min	Median	Max
Engagement Score	2,685	0.48	0.15	0.00	0.49	1.00
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

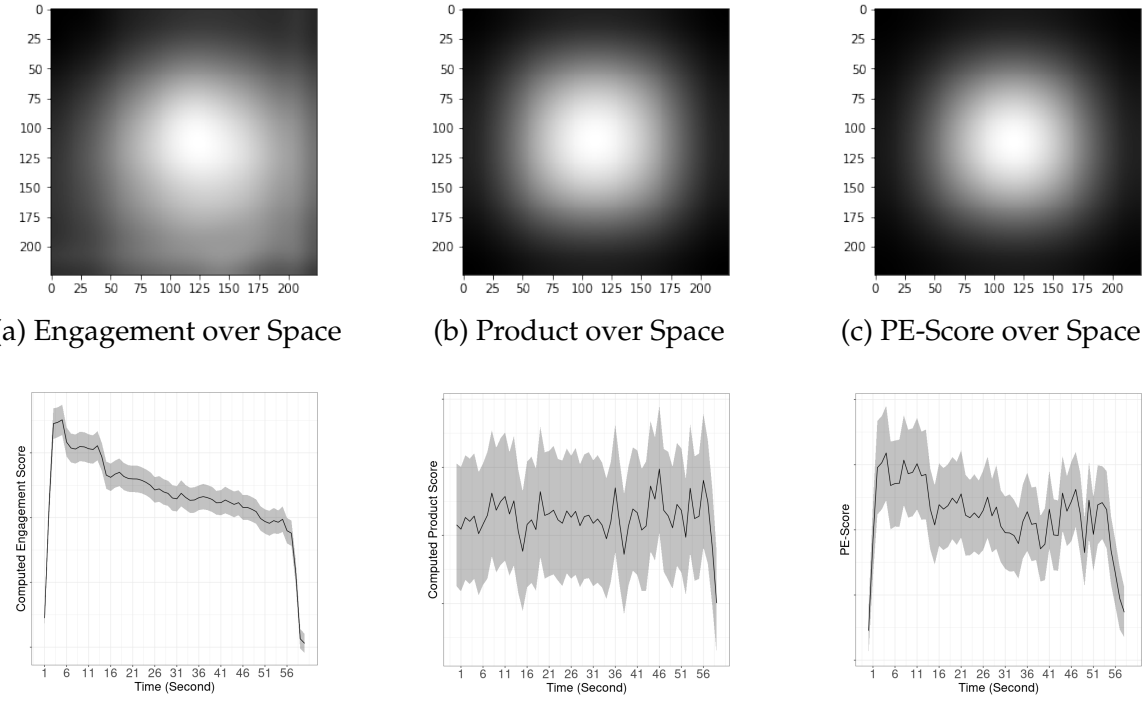
Note: The sample consists of all videos in the sales panel. All three computed scores are at the video level and normalized to [0,1]. Video-level computed engagement scores and product scores are aggregated from their pixel-level values.

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

On average, 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.[34] More importantly, 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.

[34]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).

Figure 2. Distribution of the 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 deviations from the mean.

Over the duration of a video, engagement tends to start low, 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. The 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 deviations 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-score to present suggestive evidence of incentive misalignment, the argument underlying our algorithm.

5.2 Incentive Misalignment

Our algorithm is based on the argument of influencer incentive misalignment. At the time of this study, influencers were typically paid a fixed price per video ad which was mostly driven by the number of followers and engagement (Li et al. 2023 also noted influencers' primary goal as engaging their audience). Influencers may thus have more incentive 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 interpretations 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 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 advertised 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.[35] We pool these two types of videos and regress the PE-score on an indicator variable of whether the influencer is advertising their own product while controlling for product and influencer characteristics. We find that PE-scores are, on average, 31% higher

[35] Analogously, Levitt and Syverson (2008) test incentive misalignment in the housing market comparing home sales by agents who sell for others versus themselves. See Villas-Boas (1994) and Wernerfelt et al. (2021) for analyses of the sharing and internalization of advertising agencies, respectively. See Pei and Mayzlin (2022) for a model in which influencers are paid to review products. Our "dual persona" angle is also related to Yalcin et al. (2020), who highlight the dual role of influencers as marketers and educators.

($p < 0.01$) when influencers are advertising for their own products.

Finally, our argument is that influencers are *able* to design effective video ads but may act differently for strategic reasons. To test this argument, we examine the effect of influencers' experiences. 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 the PE-score on measures of influencer experience, including the number of video ads the influencer has posted and the number of days since the influencer's first post, controlling for other influencer characteristics in Table 2.[36] We find no statistically significant association between the PE-score and influencer experience measures.

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 model-free evidence that video ads of higher PE-scores are more effective. We then evaluate our algorithm in two steps. First, we identify to which extent each video ad lifts sales. Second, we quantify to which extent the PE-score predicts sales lift, among other predictors. This two-step approach is analogous to the standard DID with heterogeneous treatment effects, which combines identification and prediction in one step in a linear model. By separating identification and prediction, the two-step approach allows each step to be more flexibly implemented. For example, one can use XGBoost, or any machine-learning model, to explore predictors of sales lift.[37]

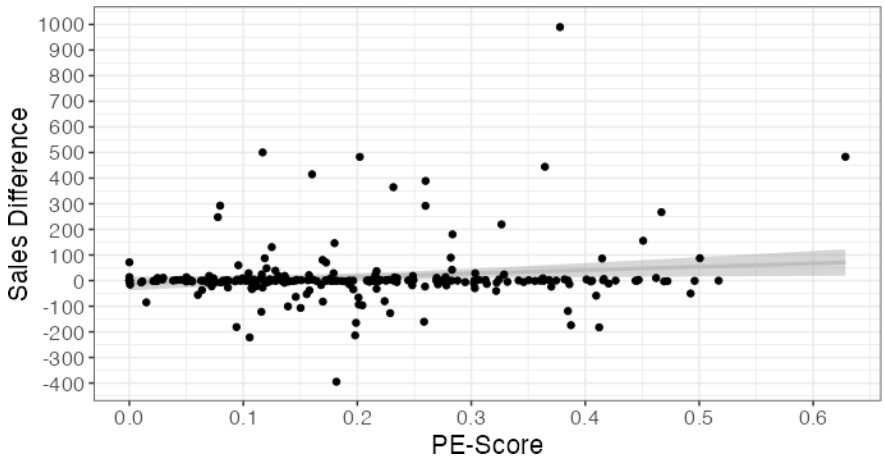
[36] There are many missing values in the number of days since the influencer's first post. To conserve the sample size, unlike other influencer characteristics in Table 2, we do not subsequently include this characteristic as a predictor of sales lift.

[37] Technically, the standard two-way fixed effect DID model has been shown to be potentially problematic when treatment adoption (ad posting in our case) is staggered (e.g., Callaway and Sant'Anna 2021). The two-step approach identifies a treatment effect (sales lift) for each video in a separate sales regression, thus avoiding the staggered-treatment problem.

5.3.1 Model-Free Evidence

We first present the model-free relationship between the PE-score and the difference in average 30-day sales revenue before and after a video ad is posted (i.e., treated). Among the 2,685 products in the sales panel, 259 were treated during the data window. We calculate each of these 259 products' sales differences as its average 30-day sales revenue within the data window after the posting of its video ad minus its average 30-day sales revenue before. We plot the sales difference against the PE-score of the corresponding video. Figure 3 shows the scatter plot. There is a positive correlation between the PE-score and sales difference ($\rho = 0.16$, $p < 0.01$), consistent with our main hypothesis. In contrast, as discussed, engagement has no significant correlation with sales difference (Figure F.5).

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



Note: This figure plots the 259 products in the sales panel that had a video ad posted (i.e., treated) during the window of the data. Each dot is a treated product. Sales difference on the y-axis equals a product's average 30-day sales revenue (in 1,000 RMB) after posting its video ad minus its average 30-day sales revenue before. We restrict the y-axis to values between -400 and 1000 for visualization and this retains over 95% 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.

5.3.2 Identifying Video-Specific Sales Lift

As the first step in our formal evaluation of the algorithm, we identify the video-level sales lift of posting an influencer video ad. We use DID for identification, leveraging the fact that different products in the sales panel posted video ads at different times (e.g., Stevenson and Wolfers 2006, Liu et al. 2019). As mentioned, among the 2,685 products in the sales panel, 259 posted video ads within the observed time window; we call them "treated products." The remaining 2,426 products posted videos before the observed time window and thus experienced no treatment event during the time window; we call them "control products." The DID method relies on the assumption that treated and control products have comparable time trends absent the treatment event. We discuss this assumption and consider alternative control products in Online Appendix H.4.

We identify each treated product's sales lift by estimating the following clustered Ordinary Least Squares (OLS) specification:

$$\textit{Daily Sales}_{vd} = \alpha \cdot \textit{Post}_{vd} + \textit{Video}_{v} + \textit{Day}_{d} + \gamma \cdot \textit{Search}_{vd} + \epsilon_{vd}. \tag{2}$$

As discussed, we focus on products that have only one video ad, so that the subscript v indexes both the video and the product. The subscript d indexes the calendar day.

The dependent variable $\textit{Daily Sales}_{vd}$ is the imputed daily sales revenue of product v on day d. As discussed, we only observe each product's 30-day sales revenue. For cleaner attribution, we impute each product's daily sales revenue from its 30-day counterpart. Drop the product subscript v for now and let $t = 1$ denote the first day a product is observed in the sales panel. For the 30 days prior, we assume $\textit{Daily Sales}_{-28} = \cdots = \textit{Daily Sales}_{1} = \textit{30-Day Sales}_{1}/30$. This smoothing rule is an approximation but is based on our partner company's observation that sales tend to be stable in this market unless there are promotional events. As a robustness check, we excluded products that either posted influencer video ads or experienced unusual fluctuations in search volume (more details to follow) during the 30 days before they enter the data window. Our conclusions

are robust. Once we have initialized daily sales for $t = -28, \cdots, 1$, we can compute daily sales for each $t \geq 2$ recursively as *Daily Sales*$_t$ = *30-Day Sales*$_t$ − *30-Day Sales*$_{t-1}$ + *Daily Sales*$_{t-30}$.[38]

The treatment variable is $Post_{vd}$ which equals 1 if video ad v is posted by day d and equals 0 otherwise. Among products in the sales panel, the 259 treated products experienced a change of $Post_{vd}$ from 0 to 1 during the observation window. The remaining 2,426 control products posted videos before the observation window so their $Post_{vd}$ is always 1 (coding $Post_{vd}$ as always 0 or 1 does not affect the estimation results).

Leveraging the panel structure of the data, we include video/product fixed effects $Video_v$ to control for unobserved heterogeneity across videos/products. We also include day fixed effects Day_d to capture common time effects (e.g., trends, seasonality) on sales. We can in theory include influencer fixed effects but their magnitude will not be separately identified from video fixed effects.

Reverse causality could be a concern if an influencer posts a video ad in anticipation of sales lift (for instance, 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 shifters and include it as a control variable, $Search_{vd}$ (Online Appendix G).[39] The parameter γ captures the effect of search and is to be estimated. The error term ϵ_{vd} is clustered by video/product and day.

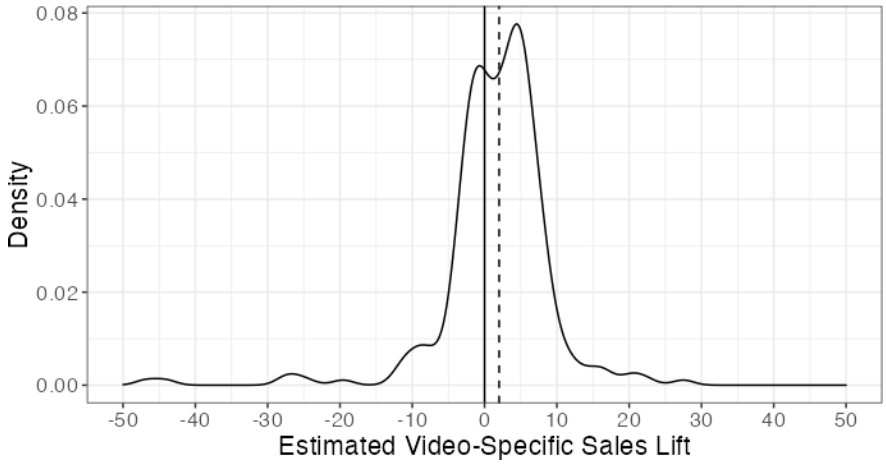
[38] If a product misses its 30-day sales revenue on a given day, we replace the missing value with the product's 30-day sales on the previous observed day to conserve the sample size.

[39] 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.

The parameter of interest in this DID estimation is α, which measures the treatment effect, or sales lift from a video ad. We separately estimate the sales lift of each of the 259 treated products using the remaining 2,426 products as controls. For each treated product, we extract its corresponding record from the sales panel and combine it with the records of the 2,426 control products in the sales panel. We estimate the sales lift of this treated product from this combined panel data of $1 + 2{,}426 = 2{,}427$ products. We repeat the DID estimation 259 times to obtain 259 sets of parameter estimates.

Figure 4 shows the distribution of the 259 sales-lift estimates across videos. Sales lift is concentrated around zero with more mass on the positive side and long tails on both sides, qualitatively consistent with previous findings on the distribution of ad treatment effects on social media (e.g., Wernerfelt et al. 2022).

Figure 4. Distribution of Sales Lift across Videos



Note: Each observation is a treated product/video in the sales panel. Sales lift is in 1,000 RMB. The vertical solid line marks zero. The vertical dashed line marks the average sales lift at 2.02. Tor visualization, the range of the x-axis is restricted to -50 to 50 (about one standard deviation of the sales-lift estimates; see Table 4).

Table 4 presents the summary statistics of the coefficients from the DID estimation. The average sales lift (α) is 2,017 RMB or \$292 and is not significantly different from zero ($p = 0.50$). This result is worth noting given that sellers pay nontrivial amounts to advertise their products; influencers in our data on average charge 19,530 RMB or \$2,826 per video ad (Table 2). It will be helpful to be able to predict sales lift before investing in

an influencer video ad, an issue we will examine in the next section. Meanwhile, search intensity shows a positive association with sales (γ) across all treated products and its mean association of 0.41 is significantly different from zero ($p < 0.001$).

Table 4. Summary Statistics of DID Estimation Results

Variable	N	Mean	St. Dev.	Min	Median	Max
Sales Lift (α)	259	2.02	47.53	-123.62	2.06	699.20
Search Coefficient (γ)	259	0.41	0.01	0.39	0.41	0.49

Note: Sales lift is estimated at the product/video level and is in 1,000 RMB. Each observation is a treated product/video in the sales panel.

5.3.3 PE-Score Predicts Sales Lift

In the second step of our algorithm evaluation, we test the predictive power of the PE-score on sales lift, among other predictors. We begin with simple OLS specifications where the dependent variable is sales lift estimated from step one and the independent variables are various sets of predictors.

As column (1) of Table 5 shows, the PE-score is a positive and significant predictor of sales lift ($p < 0.01$). The PE-score alone explains 3% of the variance in sales lift. This result is consistent with our main hypothesis that influencer video ads with higher PE-scores are more effective at lifting sales.

To test whether it is product-related engagement that predicts sales lift, we check whether overall engagement *alone* or product placement *alone* would have predictive power. As column (2) of Table 5 shows, the engagement score, computed at the video level as described in Section 5.1, is an insignificant predictor. Actual engagement (the number of likes, comments, or shares) is also insignificant. This result reaffirms industry observations and our earlier correlational finding that engagement does not necessarily predict sales lift. Column (3) of Table 5 further shows that the product score, computed at the video level as described in Section 5.1, is also an insignificant predictor. In other words, simply showing the product more in the video may not predict better sales.

Table 5. Predicting Sales Lift

		Dependent Variable: Sales Lift					
		(1)	(2)	(3)	(4)	(5)	(6)
Computed Scores	PE-Score	68.27**				101.32***	100.79***
		(22.87)				(27.04)	(28.91)
	Engagement Score		13.79			2.21	−2.44
			(20.20)			(30.55)	(32.36)
	Product Score			−6.05		−106.22	−112.47
				(21.82)		(76.98)	(80.46)
	Engagement Score × Product Score				7.10	98.61	126.50
					(40.25)	(149.97)	(156.90)
Influencer Features	Gender						5.86
							(6.37)
	# Followers						0.24
							(1.49)
	Average Play						0.31
							(2.32)
	Price per Video Ad						−9.04
							(183.65)
	Expected CPM						13.04
							(43.58)
	# Video Ads Influencer Has Posted						−0.01
							(0.18)
Product Features	Average Search						−3.67
							(7.10)
	Price						0.003
							(0.01)
	Discount						−0.002
							(0.005)
	Product-Category Indicators	No	No	No	No	No	Yes
	Observations	259	259	259	259	259	259
	R^2	0.03	0.002	0.0003	0.0001	0.06	0.08
	Adjusted R^2	0.03	−0.002	−0.004	−0.004	0.04	−0.005

Note: Each observation is a treated product/video. The specification is OLS. The dependent variable is the estimated video-level sales lift in 1,000 RMB. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

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 examine the interaction term between the engagement score and the product score as an alternative to the PE-score. This term is also insignificant, as shown in column (4) of Table 5.

Column (5) of Table 5 shows the estimation result when the PE-score, the engagement

score, the product score, and the interaction between the latter two scores are simultaneously included. The PE-score remains to be the only significant predictor ($p < 0.001$) and its effect remains positive. The PE-score is also associated with noticeable improvement of both R^2 and adjusted R^2 (which considers the number of predictors). These findings together suggest that improving engagement and product placement separately may not help; it is important that these two are aligned spatiotemporally as captured by the PE-score. This result echoes Zhang et al. (2020), who found that the temporal synchronicity between user-comment volume and movie content predicts movie enjoyment.

Last, we build on the specification in column (5) and add a rich set of covariates, including the product's average Baidu search index over the observation window, product and influencer features as reported in Table 2, and a set of product-category indicators. Column (6) of Table 5 shows the results. The PE-score continues to be a positive and significant predictor of sales lift ($p < 0.001$). In fact, it is the only significant predictor among all, including the unreported product-category indicators. The inclusion of the covariates improves R^2 but hurts adjusted R^2. This result suggests that the PE-score is a better predictor of sales lift than features, such as an influencer's number of followers, which are commonly used to select influencers for advertising. We show that popular influencers and expensive influencers may not be more effective. What they do in the video ad may be more important than who they are.

One fact to notice about Table 5 is the relatively low R^2 measures in all columns, possibly due to the functional restrictions of OLS. To evaluate the predictive power of the PE-score under more flexible specifications, we use XGBoost (Chen and Guestrin 2016),[40] a popular decision-tree based predictive-modeling algorithm, to predict sales lift with the same variables as in column (6) of Table 5. We use 5-fold cross-validation, splitting the 259 treated videos into 5 folds of (almost) equal size, where 4 folds are used to train the model and the other fold is held out to test the model. Model performance is evaluated

[40] https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf.

as the test error averaged over the 5 test folds. The XGBoost model indeed fits better than the OLS model, achieving an average R^2 of 0.42 and adjusted R^2 of 0.37 on the test folds.

Table 6 shows the 10 most important features in the XGBoost model. Gain, cover, and frequency are commonly used metrics to evaluate feature importance in decision trees. Gain is the average improvement in accuracy after a node is split based on a given feature. Cover is the fraction of the sample for which a given feature affects its decision. Frequency is the fraction of times a given feature is used to split a node. Across all three metrics, the PE-score is the most important feature that predicts sales lift.

Table 6. XGBoost Feature Importance in Predicting Sales Lift

Feature	Gain	Cover	Frequency
PE-Score	0.858	0.228	0.289
Expected CPM	0.039	0.077	0.076
Product Score	0.033	0.198	0.141
Discount	0.018	0.091	0.096
Engagement Score	0.017	0.004	0.048
# Followers	0.014	0.095	0.079
Average Search	0.009	0.101	0.069
# Video Ads Influencer Has Posted	0.003	0.062	0.038
Price	0.003	0.018	0.048
Product Category: Electronics	0.002	0.015	0.014

Note: The table reports the 10 most important features of the XGBoost model trained on the entire sample ranked by gain.

We further conduct extensive robustness analyses. Online Appendix H presents the details. In summary, our results are robust with respect to alternative construction (H.1) and validation (H.2) of the engagement heatmap, alternative definitions of engagement as predictors of sales lift (H.3), and alternative definitions of control products for causal identification of sales lift (H.4).

6 Exploring Drivers of the PE-Score

We have seen that the PE-score varies across videos and this variation matters in predicting sales lift. Ideally, we want to go beyond predictive analysis to offer prescriptive

insight on what substantive measures an influencer can take to improve the PE-score. Recall that the 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.

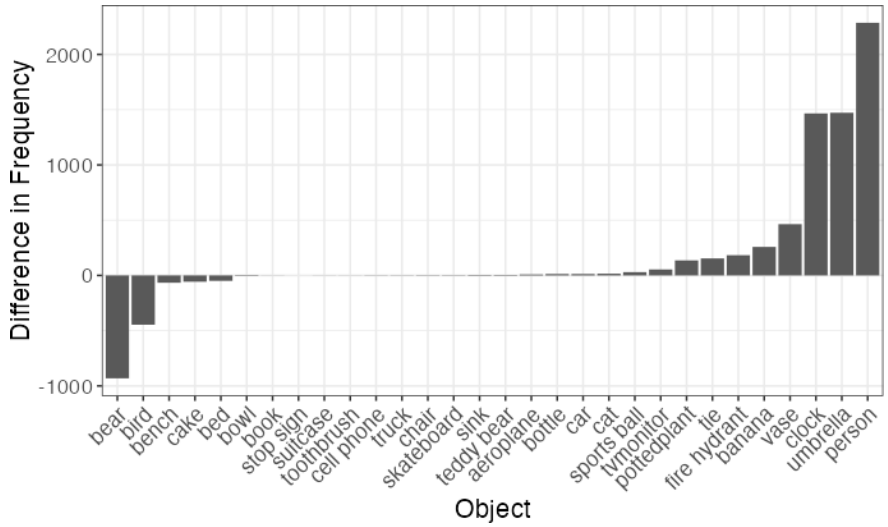
An established approach would be to use proven theories to guide the interpretation of unstructured data. For example, Zhang et al. (2020) used film grammar to analyze movie content and Zhang et al. (2022) used photography theory to evaluate image quality. The challenge in our setting, as confirmed by 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, from micro to macro, at the pixel, frame, and video-segment levels, respectively.

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),[41] 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 5 presents the results. More object instances are detected in high-engagement pixels (9,916) than in 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 ques-

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

Figure 5. Objects in High-Engagement versus Low-Engagement Pixels



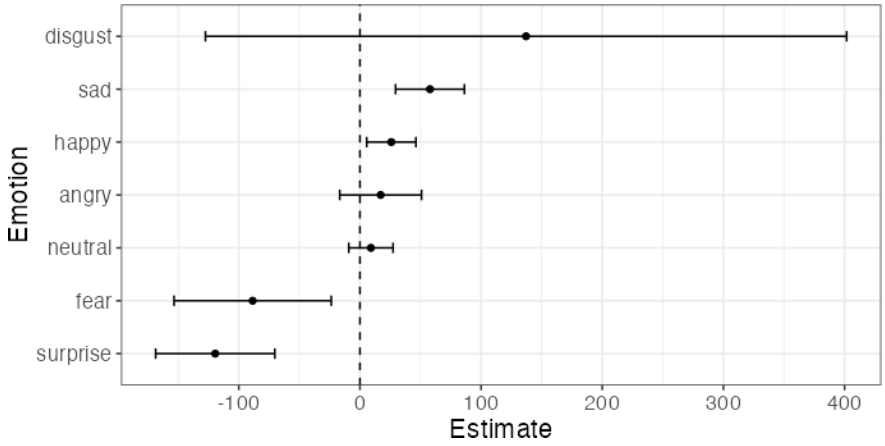
Note: Frequency is the number of times an object is detected in either the high or low version of videos in the sales panel. The difference in frequency is the frequency in high versions minus that in low versions. Some objects are only detected in high or low versions alone so that a difference cannot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

tion is what humans can do in the video to engage. Past research has identified the human face as an engaging object that attracts likes and comments on social media (Bakhshi et al. 2014, Li and Xie 2020, Hartmann et al. 2021). 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 H.2). We further ask what facial expressions drive engagement – facial expressions are arguably more actionable than factors such as facial attractiveness. We run an emotion detection algorithm, FER (Zhang et al. 2016, Arriaga et al. 2017),[42] 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 6. 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

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

that happiness and sadness are particularly contagious, taking effect in as short as half a second. The fast speed to engagement can be particularly helpful in short-form videos.

Figure 6. 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 denote the 95% confidence intervals.

Last, to expand the space of actionable recommendations beyond facial expressions, we ask what actions an influencer can take in the video to engage. 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 last 15 seconds. We run an activity detection algorithm, I3D (Carreira and Zisserman 2017),[43] on each segment to classify it into 400 pre-specified classes.[44] 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 I.1 of Online Appendix I. We can see that positive activities tend to be relatively more energetic and faster-paced (e.g., side kick, salsa dancing, krumping) or novel (e.g., getting a tattoo,

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

[44]Shorter segments may capture finer dynamics in a video but give the algorithm fewer 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., weaving basket, carving pumpkin) or mundane (e.g., garbage collecting, reading book, setting table).

We take two approaches to validate this hypothesis. First, we use topic modeling to uncover any underlying themes among these activities. Table I.1 in Online Appendix I lists the top 10 words in two topics for activities positively or negatively associated with engagement, respectively.[45] Indeed, words with higher energy, pace, or novelty (e.g., play, dance, push, climb, basketball) are identified in more-engaging activities. Words with lower energy, pace, or are more mundane (e.g., paper, question, table, wax, book) are identified in less-engaging activities.

Second, we conduct a survey with 104 college students and staff members at a university in Beijing, China 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 7. 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, quiet, 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 the 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 a moment of high emotional connectivity, or when an influencer is performing a stimulating activity. These recommendations echo

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

Figure 7. Word Clouds from the Survey of Perceived Activity Characteristics

More Engaging Activities

Less Engaging Activities

Akpinar and Berger (2017)'s finding that it is possible, and advisable, to integrate the product into viral elements of an ad for better conversion; they cited Blendtec's "Will It Blend" campaign as a good example, where the blender famously shredded an iPhone. In our framework, shredding an iPhone is engaging and the advertised product, the blender, is integral to the engagement, making it effective engagement.

7 Concluding Remarks

In this paper, we propose an algorithm to compute a metric called the PE-score, which draws on unstructured video data to predict the effect of influencer video ads on product sales. The PE-score has 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 dataset of TikTok influencer video ads and their corresponding product sales. The PE-score is a highly significant predictor of the amount of sales revenue a video ad is able to lift. The PE-score much outperforms engagement, product placement, influencer features, and product features in predicting sales lift. In fact, it is the only significant predictor among them all in the OLS framework.

Our findings are robust with respect to different ways to construct the algorithm and to identify sales lift. We also present evidence that incentive misalignment between influencers and sellers may explain the variation in the PE-score. Last, we find that engagement increases with human presence, sad or happy emotions, and stimulating or novel activities. It may be effective to integrate the product with these engaging elements.

A practical advantage of the 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 can be used to evaluate a large number of candidate videos quickly, which helps with scalability. The 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 the PE-score. Influencers can use the PE-score to aid video content development. They can make a video more engaging leveraging the actionable drivers behind the PE-score, place the product in the engaging pixels, and check the resulting PE-score for real-time feedback. Sellers can use the 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, the PE-score can serve as a metric to help clarify the attribution of sales outcomes between sellers and influencers. Finally, entertainment-commerce platforms can use the PE-score to launch various features to improve transaction efficiency. For example, a platform can highlight the PE-score as a key performance index of influencers. Providing the 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 entertainment commerce. Second, the PE-score is learned mainly through the visual components of a video ad while controlling for the acoustic features and spoken content. While a similar PE-score can be computed solely based on the sound, how to better integrate the two in an interpretable way is a worthy question.[46] Third, our exploration of engagement drivers is preliminary. Further studies including controlled experimentation may enrich the insight. Last, it will be meaningful to explore the generalizability of the PE-score. We validated the algorithm in the context of influencer video ads, where the PE-score fundamentally matters because it captures the importance of attention in entertainment commerce and because influencers may want to draw attention to themselves. However, the general principle of making product placement engaging should extend to other forms of advertising. It will be encouraging if the algorithm is able to predict sales lift based on mere ad content, being that a TV ad or a livestream marketing session (e.g., Liu et al. 2023).

References

Akpinar, E. and J. Berger (2017). "Valuable virality". *Journal of Marketing Research* 54.2, 318–330.

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.

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.

[46] A simple way to compute the PE-score for sounds is to estimate the variation of engagement over time and detect product mentions.

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.

Bentley, K., C. Chu, C. Nistor, E. Pehlivan, and T. Yalcin (2021). "Social media engagement for global influencers". *Journal of Global Marketing* 34.3, 205–219.

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.

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*, 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.

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.

Dzyabura, D., S. El Kihal, J. Hauser, and M. Ibragimov (2023). "Leveraging the power of images in predicting product return rates". *Marketing Science*, forthcoming.

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.

Fong, H., V. Kumar, and K. Sudhir (2021). "A theory-based interpretable deep learning architecture for music emotion". *SSRN 4025386*.
Gong, S., J. Zhang, P. Zhao, and X. Jiang (2017). "Tweeting as a marketing tool: a field experiment in the TV industry". *Journal of Marketing Research* 54.6, 833–850.
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.
Huang, Y. and I. Morozov (2022). "Video advertising by Twitch influencers". *SSRN 4065064*.
Itti, L. (2005). "Models of bottom-up attention and saliency". *Neurobiology of attention*, 576–582.
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.
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.
Leung, F. F., F. F. Gu, Y. Li, J. Z. Zhang, and R. W. Palmatier (2022). "Influencer marketing effectiveness". *Journal of marketing* 86.6, 93–115.
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, N., A. Haviv, and M. J. Lovett (2023). "Let's play fair – purchase and usage effects of influencer marketing on YouTube". *Available at SSRN 3884038*.
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.
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, X., D. Lee, and K. Srinivasan (2019). "Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning". *Journal of Marketing Research* 56.6, 918–943.
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.
Liu, Z., W. Zhang, X. Liu, E. Muller, and F. Xiong (2023). "Success and survival in livestream shopping". *SSRN 4028092*.
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 (2023). "Video influencers: unboxing the mystique". *SSRN 3752107*.

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*, 779–788.

Sala-i-Martin, X. X. (1997). "I just ran four million regressions". *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.

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* (15), 1929–1958.

Stevenson, B. and J. Wolfers (2006). "Bargaining in the shadow of the law: divorce laws and family distress". *Quarterly Journal of Economics* 121.1, 267–288.

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.

Tellis, G. J., D. J. MacInnis, S. Tirunillai, and Y. Zhang (2019). "What drives virality (sharing) of online digital content? The critical role of information, emotion, and brand prominence". *Journal of Marketing* 83.4, 1–20.

Testwuide, T. (2020). "Why marketing attribution has failed in the boardroom". *Forbes* (October 13).

Tian, Z., R. Dew, and R. Iyengar (2022). "Mega or micro? Influencer selection using follower elasticity". *SSRN 4173421*.

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.

Troncoso, I. and L. Luo (2022). "Look the part? The role of profile pictures in online labor markets". *Marketing Science*, forthcoming.

Tucker, C. E. (2015). "The reach and persuasiveness of viral video ads". *Marketing Science* 34.2, 281–296.

Valsesia, F., D. Proserpio, and J. C. Nunes (2020). "The positive effect of not following others on social media". *Journal of Marketing Research* 57.6, 1152–1168.

Villas-Boas, J. M. (1994). "Sleeping with the enemy: should competitors share the same advertising agency?" *Marketing Science* 13.2, 190–202.

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.

Wernerfelt, N., A. Tuchman, B. Shapiro, and R. Moakler (2022). "Estimating the value of offsite data to advertisers on Meta". *University of Chicago, Becker Friedman Institute for Economics Working Paper* 114.

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.

Yalcin, T., C. Nistor, and E. Pehlivan (2020). "Sustainability influencers: between marketers and educators". *Business Forum* 28.1.

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 (2023). "Can consumer-posted photos serve as a leading indicator of restaurant survival? Evidence from Yelp". *Management Science* 69.1, 25–50.

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.-s. and S.-c. 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 (2022). "What makes a good image? Airbnb demand analytics leveraging interpretable image features". *Management Science* 68.8, 5644–5666.

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 number 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.

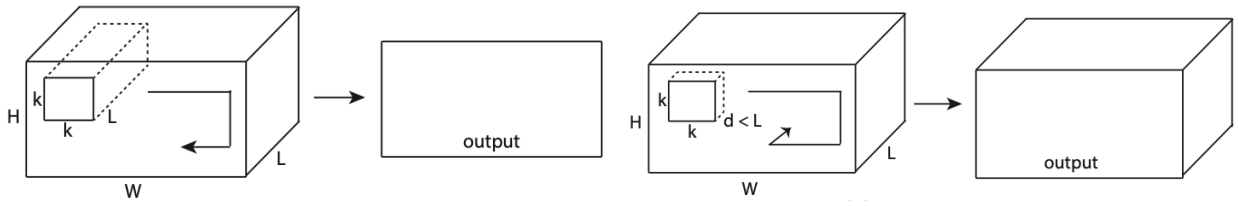
The key building blocks of a CNN are convolution layers. A convolution layer uses filters with weights that are trainable to transform the input images by representing them in a more abstract feature space that captures more general properties of the images (e.g., the presence of an edge or face). What 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 non-linear transformations of input values), and dropout layers (to randomly set some weights to zero to avoid overfitting). After layers of transformation, the feature maps are flattened into a vector and fed into a fully connected layer for the final classification or regression task.

CNNs have been used to analyze images for marketing research in a growing number of papers (e.g., Liu et al. 2020, Tkachenko and Jedidi 2020, Hartmann et al. 2021, Troncoso and Luo 2022, Zhang et al. 2022, Dzyabura et al. 2023, Zhang and Luo 2023). These papers are built upon 2D CNNs. We refer interested readers to "A Comprehensive Guide to Convolutional Neural Networks – The ELI5 Way" (Saha 2018) for a visual introduction

that animates what each layer does.[1]

In our paper, we use a 3D CNN to account for the additional time dimension of video content. We highlight the difference between a standard 2D convolution and a 3D convolution in Figure A.1. In a 2D convolution, the filter 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 matrix.

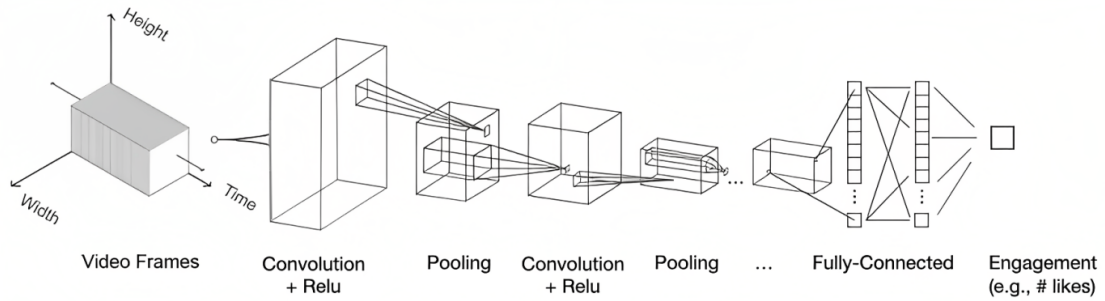
Figure A.1. 2D versus 3D Convolution (Tran et al. 2015)



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 based on 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 (TimeDistri	(None, None, 7, 7, 2048)	20861480
conv3d (Conv3D)	(None, None, 5, 5, 128)	7078016
global_max_pooling3d (Global	(None, 128)	0
dense (Dense)	(None, 256)	33024
dropout (Dropout)	(None, 256)	0
dense_1 (Dense)	(None, 1)	257

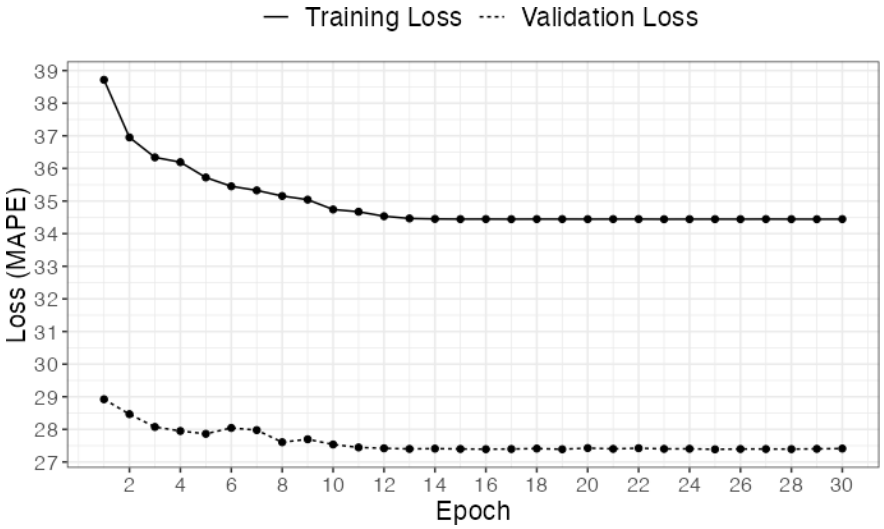
Total params: 27,972,777
Trainable params: 7,111,297
Non-trainable params: 20,861,480

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

B Engagement Heatmap

We compute the engagement heatmap as a saliency map. A saliency map is a gradient-based visualization method for CNNs (Simonyan et al. 2013). It takes a trained network and computes the gradient of the outcome with respect to a given input image. Each entry of the map represents the partial derivative of the outcome with respect to a 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). 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:

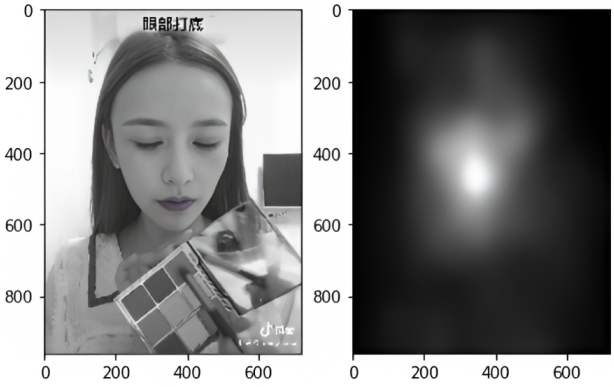
$$e_{hws} := \max_{\{r,g,b\}} \left(\left| \frac{\partial \hat{f}}{\partial x_{hwsr}} \right|, \left| \frac{\partial \hat{f}}{\partial x_{hwsg}} \right|, \left| \frac{\partial \hat{f}}{\partial x_{hwsb}} \right| \right)$$

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.

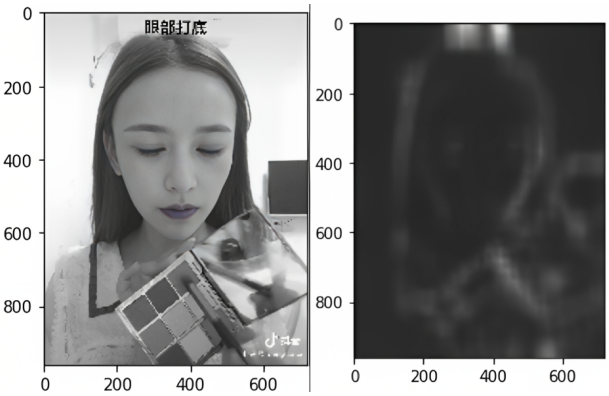
We use a saliency map to compute pixel-level engagement and call it the (supervised) engagement heatmap. It is supervised because the saliency map builds on a 3D CNN trained on video-level engagement data. In Online Appendix H.1, we also discuss an unsupervised approach to engagement heatmap that only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

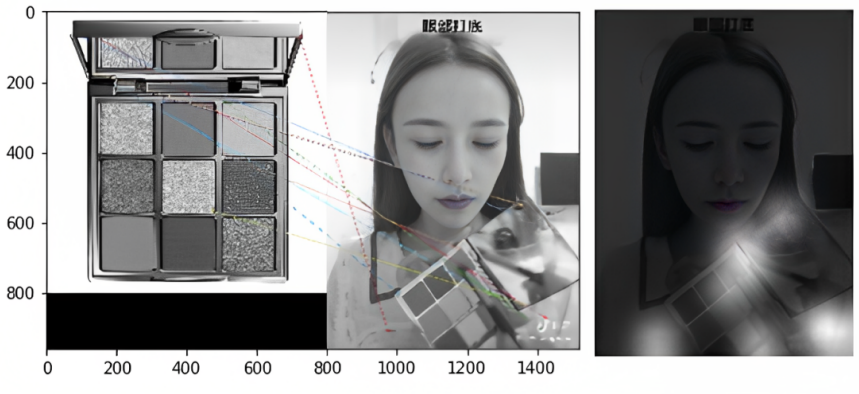
We use SIFT to detect whether an advertised product appears in a given pixel of the video. We implement SIFT via Oriented FAST and Rotated BRIEF (ORB) in OpenCV.[8]

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 ad. 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 this example, 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 the corresponding frame from the product heatmap of the video. The bright areas correspond to pixels where SIFT detects product presence. The detected product pixels can be scattered in the frame and do not necessarily enclose the entire product. We 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 with 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

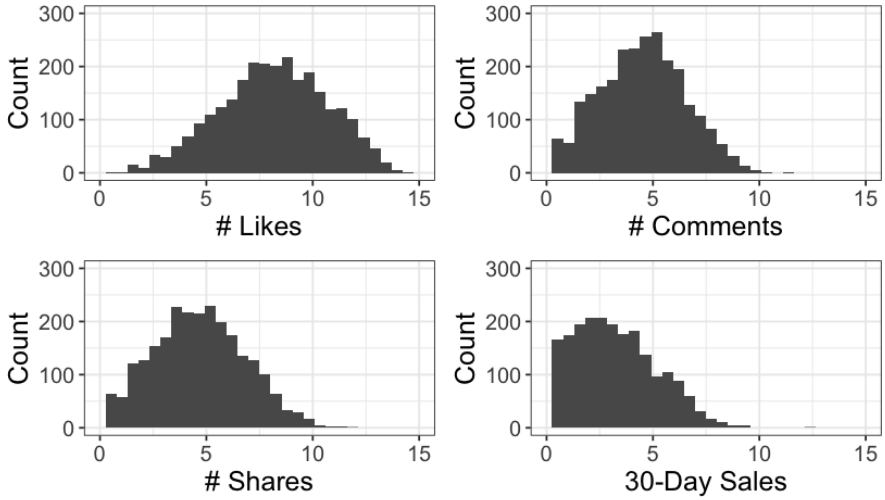
One challenge we face in our cross-category analysis is that 68% of the products in our sales panel miss category labels. Our solution is to predict missing category labels based on product titles. To do so, we draw on a sample of 8,447 products with category labels (including products outside the sales panel to increase the 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.

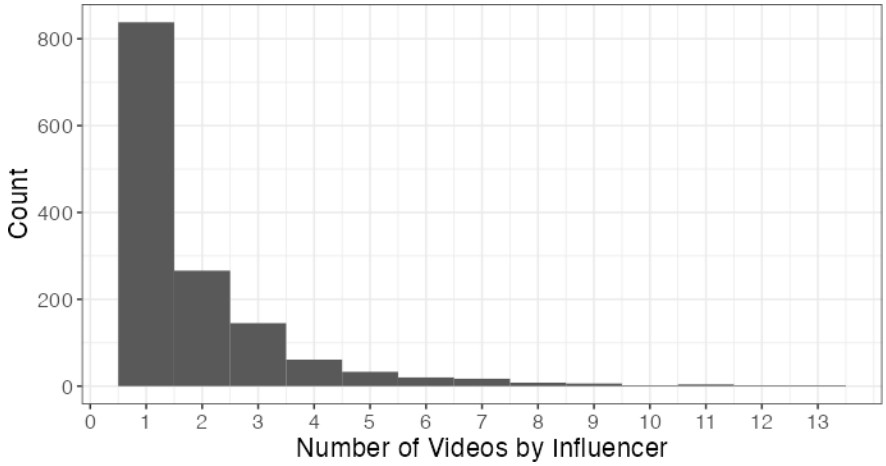
F Additional Summary Statistics of the Sales Panel

Figure F.1. Distribution of Observed Video Engagement and Product Sales



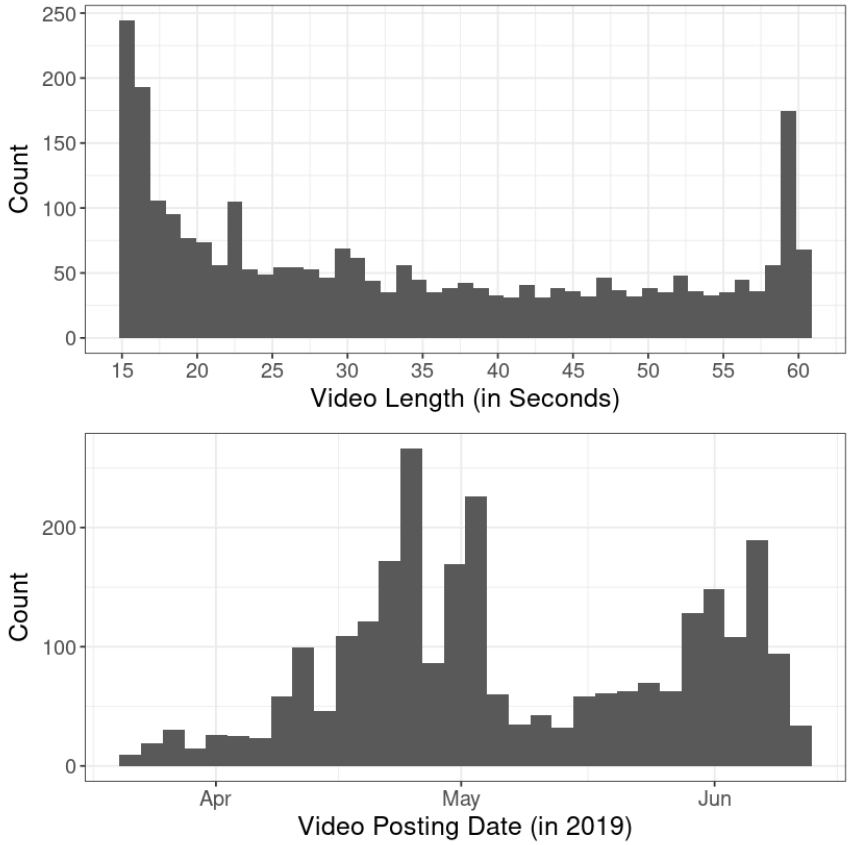
Note: The sample consists of all videos/products in the sales panel, where each product corresponds to one video ad. The subfigures present, in order, the distribution of the video-level number of likes, comments, and shares, and product-level average 30-day sales revenue as defined in the paper, all in log scale to facilitate visualization.

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



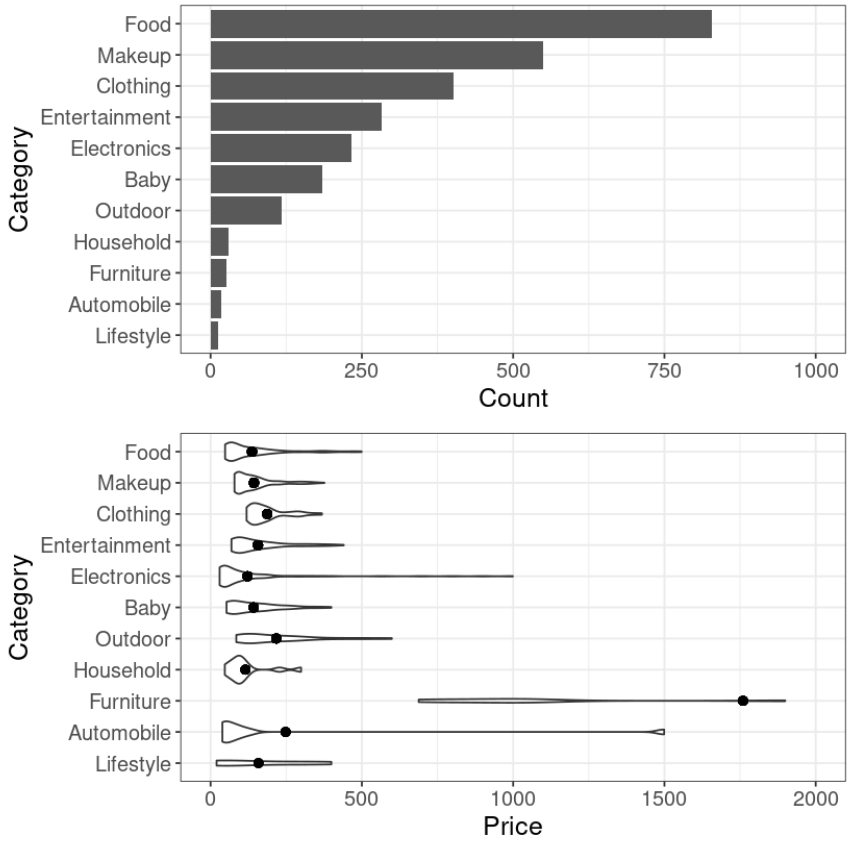
Note: The sample consists of all videos and influencers in the sales panel. Each observation is an influencer.

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



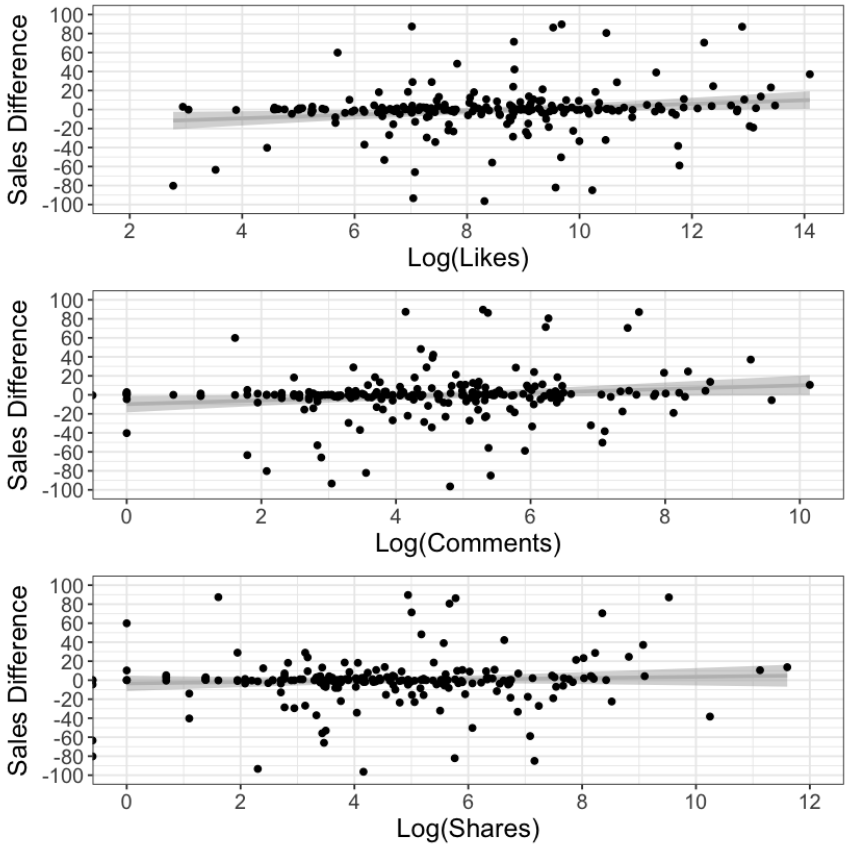
Note: The sample consists of all videos in the sales panel. Each observation is a video.

Figure F.4. Distribution of Product Categories and the Price Range



Note: The sample consists of all products in the sales panel, except that the bottom subfigure excludes products with the highest and lowest 5% prices in each category for visualization. The bottom subfigure presents a violin plot of price distributions by category. Dots represent mean prices in a category. Prices are in RMB.

Figure F.5. Before-After Sales Difference by Engagement



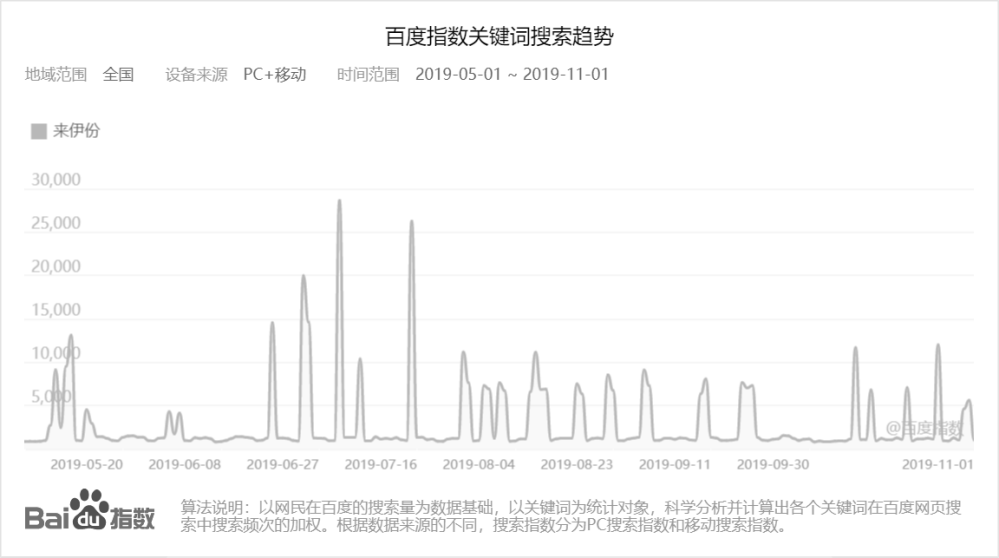
Note: Sales difference on the y-axis equals a product's average 30-day sales revenue (in 1,000 RMB) after posting its video ad minus its average 30-day sales revenue before. The x-axis is the observed video-level engagement metric (the number of likes, comments, and shares) on a logarithmic scale to control for outliers. Each dot is a treated product as defined in the paper. We restrict the y-axis to values between -100 and 100 for visualization. Sales difference has no significant correlation with any of the observed engagement metrics in the full sample ($\rho = -0.0075$, $p = 0.90$ for likes; $\rho = -0.02$, $p = 0.74$ for comments; $\rho = -0.0074$, $p = 0.91$ for shares). Gray areas represent the 95% confidence band along the regression line.

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 the 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 searches 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 a 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 for 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 the identification of sales lift because we include product fixed effects in the analysis.

H Robustness Checks and Extensions

In this section, we extend the main analysis presented in the paper 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.

For all robustness checks, we test the predictive power of the PE-score in both the OLS and XGBoost frameworks. We report robustness-check results when we rerun column (6) of Table 5 (the OLS specification with the most features) and Table 6 (the table of the most important features in XGBoost).

H.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 extract saliency maps. As a first robustness check, we retrain the algorithm using the number of likes and comments instead and rerun subsequent analysis that relies on the PE-score. (The identification of sales lift is not affected.) Columns (1) and (2) of Table H.1 present the OLS estimation results. The PE-score remains a positive predictor and the only significant predictor of sales lift ($p < 0.05$). Tables H.2 and H.3 present the most important features from the XGBoost model. The PE-score continues to be the most important predictor of sales lift.

So far, we have followed the supervised approach to construct the engagement heatmap, using video content as input and observed video-level engagement (the number of shares, likes, or comments) as output. Pixel-level engagement is thus determined in a supervised way; a pixel will have a high engagement score if a small change in its value affects observed video-level engagement by a large amount. We next check if the algorithm works when engagement is computed in an unsupervised approach.

Table H.1. Predicting Sales Lift
(Alternative Construction of the Engagement Heatmap)

		Dependent Variable: Sales Lift		
		(1) Constructed on # Likes	(2) Constructed on # Comments	(3) Constructed Unsupervised
Computed Scores	PE Score	119.03*	160.33*	24.69
		(53.35)	(67.34)	(30.23)
	Engagement Score	−5.84	−0.99	−10.67
		(32.87)	(33.07)	(33.21)
	Product Score	−78.16	−75.53	−92.45
		(81.09)	(80.99)	(83.36)
	Engagement Score × Product Score	135.89	130.49	162.86
		(159.36)	(159.26)	(160.28)
Influencer Features	Gender	6.55	7.56	8.04
		(6.47)	(6.42)	(6.49)
	# Followers	0.69	0.54	0.64
		(1.51)	(1.51)	(1.53)
	Average Play	0.62	0.40	0.19
		(2.36)	(2.35)	(2.38)
	Price per Video Ad	−68.89	−69.09	−69.93
		(185.09)	(184.79)	(188.29)
	Expected CPM	10.31	8.73	5.70
		(44.49)	(44.40)	(44.84)
	# Video Ads Influencer Has Posted	−0.01	0.01	−0.01
		(0.18)	(0.18)	(0.18)
Product Features	Average Search	−3.60	−4.83	−4.77
		(7.22)	(7.19)	(7.27)
	Price	0.002	0.002	0.002
		(0.01)	(0.01)	(0.01)
	Discount	−0.003	−0.003	−0.003
		(0.005)	(0.005)	(0.005)
	Product-Category Indicators	Yes	Yes	Yes
	Observations	259	259	259
	R^2	0.05	0.06	0.04
	Adjusted R^2	−0.03	−0.03	−0.05

Note: Each observation is a treated product/video. The specification is OLS. The dependent variable is the estimated video-level sales lift in 1,000 RMB. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table H.2. XGBoost Feature Importance in Predicting Sales Lift
(Engagement Heatmap Constructed on # Likes)

Feature	Gain	Cover	Frequency
PE-Score	0.864	0.257	0.303
Discount	0.037	0.140	0.092
Product Score	0.031	0.164	0.209
Engagement Score	0.026	0.008	0.041
# Followers	0.015	0.059	0.062
Expected CPM	0.008	0.095	0.073
Average Search	0.006	0.024	0.026
Average Play	0.003	0.054	0.051
# Video Ads Influencer Has Posted	0.002	0.051	0.053
Price per Video Ad	0.002	0.020	0.019

Note: The table reports the 10 most important features of the XGBoost model trained on the entire sample ranked by gain.

Table H.3. XGBoost Feature Importance in Predicting Sales Lift
(Engagement Heatmap Constructed on # Comments)

Feature	Gain	Cover	Frequency
PE-Score	0.857	0.257	0.292
Product Score	0.041	0.209	0.149
Engagement Score	0.021	0.052	0.104
Discount	0.020	0.091	0.070
Expected CPM	0.019	0.089	0.077
Average Search	0.012	0.059	0.038
# Video Ads Influencer Has Posted	0.011	0.049	0.043
# Followers	0.011	0.064	0.063
Product Category: Electronics	0.003	0.011	0.018
Average Play	0.002	0.044	0.050

Note: The table reports the 10 most important features of the XGBoost model trained on the entire sample ranked by gain.

The motivation for 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 by the images themselves.

To construct the unsupervised engagement heatmap, we use the intrinsic properties of the images (the statistically distinct areas of an image, such as high contrast locations and edges of objects; see Figure B.1b for an example) as a proxy for pixel-level engagement (Hou and Zhang 2007). As column (3) of Table H.1 shows, the PE-score based on unsupervised learning does not predict sales lift with statistical significance. As Table H.4 shows, the PE-score is no longer the most important predictor of sales lift based on two out of three importance metrics. These results suggest that statistically distinct areas do not imply higher engagement. There is substantial value in collecting engagement data to construct a supervised engagement heatmap.

Table H.4. XGBoost Feature Importance in Predicting Sales Lift
(Unsupervised Engagement Heatmap)

Feature	Gain	Cover	Frequency
Average Play	0.768	0.140	0.098
Expected CPM	0.085	0.176	0.097
PE-Score	0.031	0.134	0.218
Product Score	0.030	0.145	0.128
Price per Video Ad	0.029	0.052	0.057
Engagement Score	0.021	0.041	0.088
Discount	0.009	0.071	0.093
Product Category: Clothing	0.006	0.031	0.021
# Followers	0.006	0.068	0.057
Average Search	0.004	0.018	0.028

Note: The table reports the 10 most important features of the XGBoost model trained on the entire sample ranked by gain.

H.2 Validity Check of the Engagement Heatmap

As discussed in the paper, the literature has identified the human face as an engaging object that attracts likes and comments on social media (Bakhshi et al. 2014, Li and Xie 2020, Hartmann et al. 2021). 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.[15] 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 correlation is indeed positive and significant ($\rho = 0.04$, $p < 0.001$). 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.

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

H.3 Observed Engagement Measures as Predictors of Sales Lift

The main analysis in the paper uses each video's computed engagement score as a predictor of its sales lift. We check the robustness of our results if we replace a video's computed engagement score with observed engagement measures – the number of likes, comments, or shares. Table H.5 presents the OLS estimation results, where the PE-score continues to a positive predictor and the only significant predictor of sales lift ($p < 0.001$).

Table H.5. Predicting Sales Lift
(Observed Engagement Measures as Predictors)

		Sales Lift		
		(1) Engagement as # Likes	(2) Engagement as # Comments	(3) Engagement as # Shares
Computed Scores & Observed Engagement	PE Score	100.62***	100.98***	100.81***
		(28.84)	(28.96)	(28.80)
	Observed Engagement	24.76	603.39	732.33
		(72.52)	(3,199.10)	(969.85)
	Product Score	−45.94	−47.63	−46.08
		(29.30)	(28.57)	(27.83)
	Observed Engagement × Product Score	−107.71	−2,719.43	−5,366.47
		(290.16)	(8,948.66)	(6,426.91)
Influencer Features	Gender	5.32	5.20	5.31
		(6.41)	(6.36)	(6.35)
	# Followers	−0.02	−0.09	−0.62
		(1.61)	(2.13)	(1.88)
	Average Play	0.22	0.32	0.63
		(2.34)	(2.38)	(2.37)
	Price per Video Ad	−1.02	2.50	30.73
		(204.01)	(193.85)	(193.00)
	Expected CPM	12.98	12.67	11.64
		(43.97)	(43.97)	(43.92)
	# Video Ads Influencer Has Posted	−0.01	−0.01	−0.01
		(0.18)	(0.18)	(0.18)
Product Features	Average Search	−3.35	−3.37	−3.65
		(7.12)	(7.12)	(7.12)
	Price	0.003	0.003	0.002
		(0.01)	(0.01)	(0.01)
	Discount	−0.002	−0.002	−0.002
		(0.005)	(0.005)	(0.005)
	Product-Category Indicators	Yes	Yes	Yes
	Observations	259	259	259
	R^2	0.08	0.08	0.08
	Adjusted R^2	−0.01	−0.01	−0.01

Note: Each observation is a treated product/video. The specification is OLS. The dependent variable is the estimated video-level sales lift in 1,000 RMB. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

We rerun XGBoost replacing the computed engagement score with observed engagement. For a more succinct and strict test, we include all three measures of observed engagement (the number of likes, comments, and shares) simultaneously in the XGBoost model. The PE-score continues to be the most important predictor of sales lift.

Table H.6. XGBoost Feature Importance in Predicting Sales Lift
(Observed Engagement Measures as Predictors)

Feature	Gain	Cover	Frequency
PE-Score	0.861	0.230	0.287
Expected CPM	0.037	0.081	0.068
Product Score	0.036	0.149	0.135
# Comments	0.016	0.099	0.070
Discount	0.013	0.064	0.068
# Likes	0.010	0.056	0.073
Average Search	0.007	0.051	0.028
# Shares	0.006	0.063	0.070
Price per Video Ad	0.005	0.023	0.031
# Followers	0.003	0.060	0.054

Note: The table reports the 10 most important features of the XGBoost model trained on the entire sample ranked by gain.

H.4 Alternative Definitions of Control Products

We revisit our causal identification of sales lift in this section. When the treatment effect varies over time, it might be problematic to use already-treated products as controls (Callaway and Sant'Anna 2021). We address this potential concern with two alternative ways to define control products.

First, we only include in the control group products that did not post influencer video ads or experience unusual fluctuations in search volume during the 30 days before they enter the data window – any impact of the video ad on the product's daily sales trajectory is likely to have diminished given that video popularity tends to be short-lived on TikTok. For context, Huang and Morozov (2022) analyzed video advertising by Twitch influencers and found an hourly ad carryover coefficient of 0.828, which is equivalent to a daily carryover coefficient of only 0.011. Excluding products that posted video ads or experienced unusual fluctuations in search volume within the prior-30-day window leaves us with 321 control products.

Second, we use not-yet-treated products and their pre-treatment daily sales trajectories as controls. Specifically, we only include products that posted their video ads after June 1, 2019 and their sales panel before their respective ad posting dates as controls. This splits the 259 treated products into 146 treated products and 113 control products.

Table H.9 presents the OLS-estimation results, where the pe-score remains a positive predictor and the only significant predictor of sales lift ($p < 0.001$) under both alternative definitions of control products. Tables H.10 and H.11 report the XGBoost feature importance results under the two definitions of control products, respectively. The pe-score remains the most important predictor of sales lift. Notably, the predictive power of the pe-score remains distinctively strong despite the much smaller number of control products under both alternative definitions.

Table H.7. Summary Statistics of DID Estimation Results (Excluding Products with Ads 30 Days Prior from Control Group)

Variable	N	Mean	St. Dev.	Min	Median	Max
Sales Lift (α)	259	1.56	48.01	-126.30	1.88	710.79
Search Coefficient (γ)	259	0.25	0.86	-4.09	0.19	10.99

Note: Sales lift is estimated at the product/video level and is in 1,000 RMB. Each observation is a treated product/video in the sales panel.

Table H.8. Summary Statistics of DID Estimation Results (Including Only Not-Yet-Treated Products in Control Group

Variable	N	Mean	St. Dev.	Min	Median	Max
Sales Lift (α)	146	19.73	180.81	-101.07	2.45	2,178.65
Search Coefficient (γ)	146	-29.25	2.03	-30.17	-29.53	-10.17

Note: Sales lift is estimated at the product/video level and is in 1,000 RMB. Each observation is a treated product/video in the sales panel.

Table H.9. Predicting Sales Lift
(Alternative Definitions of Control Products)

		Dependent Variable: Sales Lift	
		(1) Excluding Products with Ads 30 Days Prior from Control Group	(2) Including Only Not-Yet-Treated Products in Control Group
Computed Scores	PE-Score	103.04***	601.38***
		(29.17)	(164.31)
	Engagement Score	−2.98	168.97
		(32.65)	(182.23)
	Product Score	−110.90	−235.92
		(81.18)	(456.95)
	Engagement Score × Product Score	115.07	−183.98
		(158.29)	(880.90)
Influencer Features	Gender	5.65	0.93
		(6.42)	(33.49)
	# Followers	0.19	−4.53
		(1.50)	(6.81)
	Average Play	0.06	6.24
		(2.34)	(13.30)
	Price per Video Ad	−11.21	705.57
		(185.28)	(981.38)
	Expected CPM	15.00	27.61
		(44.24)	(171.48)
	# Video Ads Influencer Has Posted	0.01	−0.74
		(0.18)	(1.15)
Product Features	Average Search	−3.96	−9.38
		(7.17)	(36.67)
	Price	0.003	0.05
		(0.01)	(0.09)
	Discount	−0.002	−0.06
		(0.005)	(0.08)
	Product-Category Indicators	Yes	Yes
	Observations	259	146
	R^2	0.08	0.15
	Adjusted R^2	−0.002	0.0005

Note: Each observation is a treated product/video. For column (2), some of the 259 treated products are reclassified as control products. The specification is OLS. The dependent variable is the estimated video-level sales lift in 1,000 RMB. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Table H.10. XGBoost Feature Importance in Predicting Sales Lift
(Excluding Products with Ads 30 Days Prior from the Control Group)

Feature	Gain	Cover	Frequency
PE-Score	0.882	0.220	0.281
Product Score	0.035	0.249	0.169
Expected CPM	0.025	0.103	0.094
Engagement Score	0.017	0.032	0.114
Discount	0.016	0.088	0.058
# Followers	0.008	0.083	0.072
Average Search	0.004	0.095	0.050
Product Category: Electronics	0.003	0.014	0.017
# Video Ads Influencer Has Posted	0.003	0.030	0.028
Price	0.003	0.050	0.047

Note: The table reports the 10 most important features of the XGBoost model trained on the entire sample ranked by gain.

Table H.11. XGBoost Feature Importance in Predicting Sales Lift
(Including Only Not-Yet-Treated Products in the Control Group)

Feature	Gain	Cover	Frequency
PE-Score	0.991	0.208	0.287
Product Score	0.003	0.195	0.132
Discount	0.002	0.115	0.088
# Video Ads Influencer Has Posted	0.001	0.090	0.073
Price per Video Ad	0.001	0.053	0.047
Engagement Score	0.001	0.077	0.114
Product Category: Household	0.001	0.049	0.019
Average Play	0.001	0.053	0.044
# Followers	0.0004	0.062	0.079
Price	0.0002	0.057	0.054

Note: XGBoost is run on the subset of 146 treated products/videos using 5-fold cross-validation. The table reports the 10 most important features of the XGBoost model ranked by gain.

H.5 Alternative Imputation of Daily Sales

We leverage observed daily search patterns to impute daily sales in a different way. Recall that the dependent variable *Daily Sales*$_{vd}$ is the imputed daily sales revenue of product v on day d. As discussed, we only observe each product's 30-day sales revenue. For cleaner attribution, we impute each product's daily sales revenue from its 30-day counterpart. Drop the product subscript v for now and let $t = 1$ denote the first day a product is observed in the sales panel. For the 30 days prior, we assume *Daily Sales*$_t$ $=$ *30-Day Sales*$_1$ $\times \frac{Search_t}{\sum_{t'=t-29}^{0} Search_{t'}}$. This smoothing rule is an approximation based on the assumption that daily sales is likely to be correlated with daily search volume. When we do not observe at least 30 days of search volume prior to the first day a product is observed in the sales panel, we attribute sales equally for the 30 days prior (the same approach as in the main text).

Table H.12. Summary Statistics of DID Estimation Results (Alternative Imputation)

Variable	N	Mean	St. Dev.	Min	Median	Max
Sales Lift (α)	259	1.92	47.51	-123.58	2.03	699.21
Search Coefficient (γ)	259	0.30	0.01	0.28	0.30	0.39

Note: Sales lift is estimated at the product/video level and is in 1,000 RMB. Each observation is a treated product/video in the sales panel.

Table H.13. Predicting Sales Lift (Alternative Imputation) – Format to be updated

	Dependent variable:
	tau_alt
pe_score	101.08***
	(28.88)
e_score	−2.46
	(32.33)
p_score	−115.00
	(80.37)
p_score:e_score	129.74
	(156.71)
gender	5.72
	(6.36)
fans	0.13
	(1.49)
avg_play	0.45
	(2.32)
influencer_price	8.99
	(183.43)
expected_cpm	12.51
	(43.79)
order_cnt	−0.05
	(0.18)
avg_search	−3.53
	(7.10)
price	0.003
	(0.01)
discount	−0.002
	(0.005)
Observations	259
R^2	0.08
Adjusted R^2	−0.003
Note:	*p<0.1; **p<0.05; ***p<0.01

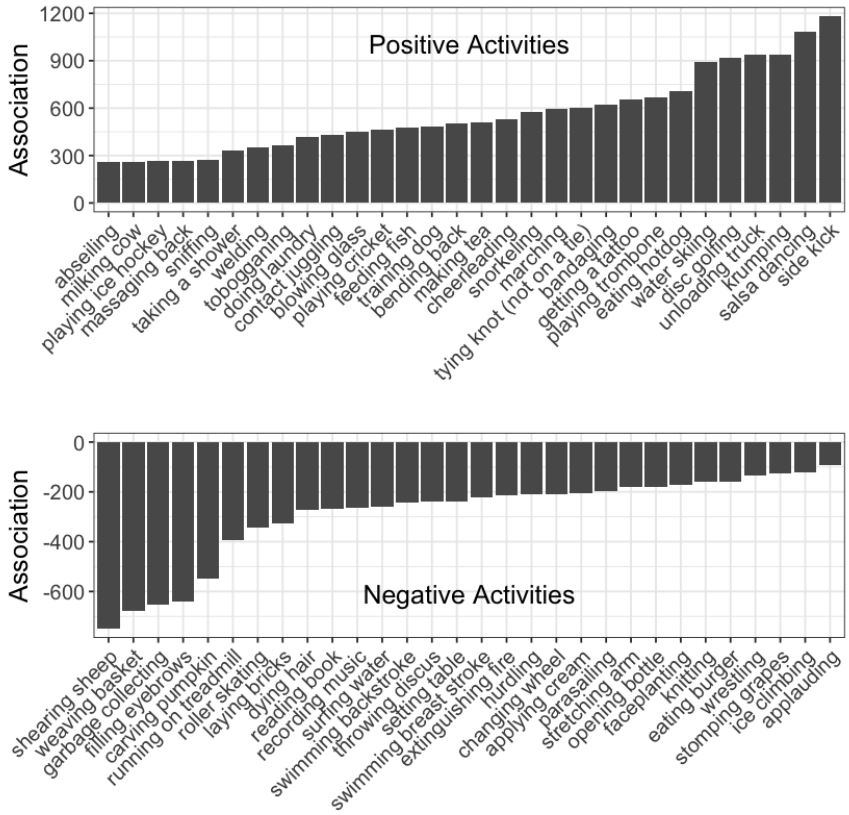
Table H.14. XGBoost Feature Importance in Predicting Sales Lift (Alternative Imputation)

Feature	Gain	Cover	Frequency
PE-Score	0.862	0.258	0.301
Expected CPM	0.041	0.063	0.060
Product Score	0.031	0.246	0.176
Engagement Score	0.018	0.023	0.081
Product Category: Food	0.009	0.011	0.019
Discount	0.007	0.078	0.079
Average Search	0.006	0.055	0.035
# Video Ads Influencer Has Posted	0.006	0.030	0.035
# Followers	0.006	0.085	0.068
Price	0.005	0.051	0.049

Note: The table reports the 10 most important features of the XGBoost model trained on the entire sample ranked by gain.

I Activities in the Video Ad and Engagement

Figure I.1. Association between Activities and Engagement Scores



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

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

Positive Activities		Negative Activities	
Topic 1	Topic 2	Topic 1	Topic 2
play	play	swim	appli
danc	make	paper	paper
make	clean	question	fill
clean	climb	tabl	swim
use	ice	fli	run
back	blow	feed	stretch
basketbal	car	die	leg
fold	eat	basket	wax
eat	floor	monopoli	facepl
push	head	bottl	book