Mega or Micro?
Influencer Selection Using Follower Elasticity

Zijun Tian, Ryan Dew, Raghuram Iyengar*
University of Pennsylvania

July 28, 2022

Abstract

Despite the explosive growth of influencer marketing, wherein companies sponsor social media personalities to promote their brands, there is little research to guide companies' selection of influencer partners. One common criterion is popularity: while some firms sponsor "mega" influencers with millions of followers, other firms partner with "micro" influencers, who may only have several thousands of followers, but may also cost less to sponsor. To quantify this trade-off between reach and cost, we develop a framework for estimating the *follower elasticity of impressions*, or FEI, which measures a video's percentage gain in impressions corresponding to a percentage increase in the follower size of its creator. Computing FEI involves estimating the causal effect of an influencer's popularity on the view counts of their videos, which we achieve through a combination of a unique dataset collected from TikTok, a representation learning model for quantifying video content, and a machine learning-based causal inference method. We find that FEI is always positive, but often nonlinearly related to follower size, suggesting different optimal sponsorship strategies than those observed in practice. We examine the factors that predict variation in these FEI curves, and show how firms can use these results to better determine influencer partnerships.

Keywords: influencer marketing, causal inference, deep learning, representation learning, heterogeneous treatment effects, video data

*Zijun Tian is a doctoral student in Economics at the University of Pennsylvania. Ryan Dew is an Assistant Professor of Marketing at the Wharton School of the University of Pennsylvania, and Raghuram Iyengar is the Miers-Busch, W'1885 Professor, Professor of Marketing at the Wharton School of the University of Pennsylvania. Zijun Tian is the corresponding author: zjtian96@sas.upenn.edu

1 Introduction

Influencer marketing has become an integral part of advertising strategy as companies recognize the importance of collaborating with social media personalities. In its most basic form, firms pay influencers to share content on social platforms with twofold goals of reaching the target audience and gaining their trust. While there are many outlets for such sponsorships (e.g., Instagram and Snap), TikTok has emerged as the most popular one.[1] TikTok is a video-focused social networking service and is largely attributed with popularizing the short form user video. With typical lengths ranging from 15- to 180-seconds, these videos are ideal for holding the attention of an audience with an ever-shrinking attention span. In the past few years, many brands have become active on TikTok by promoting *challenges*. A typical challenge contains a name in the form of a hashtag (e.g., #MakeMomSmile from Colgate), a sentence or two encouraging users to create content matching a theme (e.g., "Make Mom Smile this Mother's Day by doing something special for mothers"), and a few user-generated videos, sometimes sponsored by the company, that can jumpstart the campaign.[2]

Figure 1 shows two examples of sponsored posts from two challenges. The left panel shows a fashion influencer who created a dancing video for Dettol's #HandWashChallenge that aimed to promote its hand-washing products. The right panel shows a family influencer who participated in Walmart's #UnwrapTheDeals campaign, which featured a special effect animation which TikTok users could add to their videos to advertise Walmart's Black Friday deals. As marketers consider shifting their expenditures toward influencer marketing, there is a growing need to assess the value of sponsoring influencers to create official videos like these, which set the tone for the challenge, and are key to it gaining wide visibility. An on-going debate is the importance of influencers with large numbers of followers (e.g., so-called "mega influencers," with millions of followers) as best suited for promoting a campaign, as compared to those with smaller followings (e.g., "micro influencers," with thousands followers). While the former have a larger potential audience, they are typically more expensive than the latter.[3] Moreover, the issue is further complicated by the fact that estimating the effect of followers on an outcome like reach is fundamentally

[1] https://www.emarketer.com/content/tiktok-influencer-marketing

[2] https://www.tiktok.com/tag/MakeMomSmile

[3] https://www.businessofapps.com/marketplace/influencer-marketing/research/influencer-marketing-costs/

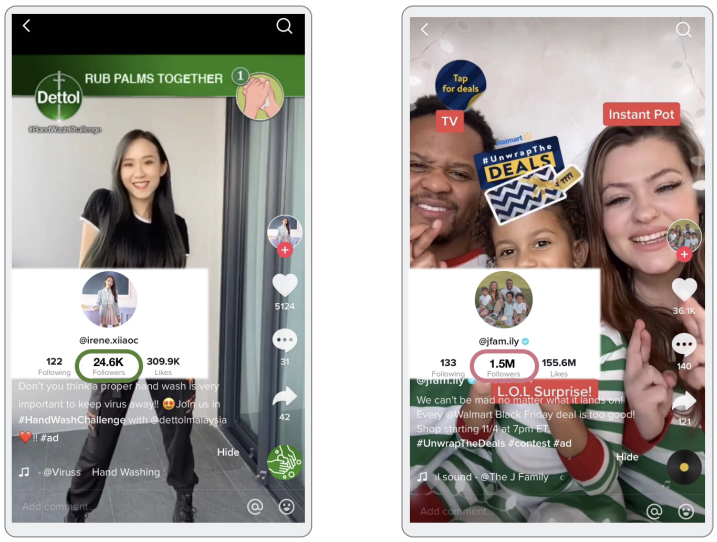


Figure 1: Examples of Sponsored Videos on TikTok
At left, a screenshot of a video posted posted by a mid-sized fashion influencer under Dettol's #HandWashChallenge. At right, a screenshot of a video posted by a mega influencer under Walmart's #UnwrapTheDeals campaign.

a *causal* question, which poses a number of challenges. We use the two sponsorships shown in Figure 1 to discuss the key issues.

First, the content of an influencer's videos affects both how many followers that influencer has, and how popular those videos become. Consider, for example, the two influencers in Figure 1: the influencer in the right panel of the figure tends to post content related to families, which may have broad appeal on TikTok, while the influencer in the left panel of the figure typically posts about fashion, which may have more niche appeal. If this is true, then without accounting for the different content, we may erroneously infer that follower size, rather than content, is driving the popularity of the videos. Statistically controlling for the content of a video, however, poses a challenge: video data is unstructured, containing different modalities (images, audio, text) that work in synergy to convey the content. Even if we are able to quantify the content of a video, there still remain a number of other challenges in measuring the causal effect of interest. Another concern is that the causal effect of follower size on video impressions might, itself, be difficult to measure effectively. It may, for instance, be nonlinear, akin to the well-known S-shaped response curve in advertising (Johansson, 1979), or it may differ by the type of content itself (i.e., there may be heterogeneity in the treatment effect). For instance, perhaps an instructional video about fashion may benefit less from each marginal follower than a social challenge. Finally, while content is

the most obvious confounder in our setting, there are a number of other potentially unobserved confounders that may further complicate estimating the effect: for instance, perhaps the family influencer at right tends to advertise their videos on another social media site, while the fashion influencer does not. These cross-posting behaviors would be difficult to measure, but may simultaneously affect how many followers an account has, and how many views their videos get, again leading to potentially erroneous inferences about the relationship between the two.

In this paper, we develop a modeling framework for determining the causal relationship between the number of followers that an influencer has and the reach of their posts that addresses all of the above issues. We do so through two key components. First, to summarize the content of a video, we employ a representation learning model based on the variational autoencoder (or VAE). This part of our model takes as its inputs all the observed information about a post (e.g., video, hashtags), and outputs a dense vector representation of the content. Second, conditional on the learned representation, and other observables about the creator and post, we leverage a recently introduced, machine learning-based causal inference framework called Deep Instrumental Variables (Hartford et al., 2017). The Deep IV approach allows for the flexible estimation of potentially nonlinear and heterogeneous relationships between cause and effect, in the case when a valid instrumental variable is present, which crucially, is the case in our empirical setting: TikTok.

We estimate our model using a curated dataset of TikTok videos that appeared on its Discover page over a period of six months. The Discover page on TikTok shows users a wide variety of videos grouped by categories that are currently trending in the TikTok community. The page is updated daily with 1-2 new hashtags, giving us a wide snapshot of popular content on TikTok. Our dataset consists of more than 200 hashtags, and just over half a million videos that are publicly accessible under those hashtags. Each of these videos, as well as their creators, were tracked daily for two weeks after the creation of each video, allowing us to record our two variables of interest: the follower count of an influencer at the time a video was created, and the impression count of a video two weeks after its posting. Finally, TikTok provides an ideal setting for our analysis, because it has revealed in public posts aspects of how its recommendation algorithm works. As we subsequently explain, this information suggests a valid instrument we can use in conjunction with the Deep IV approach.

From this modeling framework and our TikTok data, we derive a key quantity of interest for

firms determining how "big" of an influencer to sponsor: the follower elasticity of impressions, or FEI. Based on our causal analysis, the FEI captures, for a given video, the expected percentage change in that video's impressions corresponding to a unit percentage change in its creator's follower size. We quantify FEI over all possible values of the number of followers, which results in a curve that shows how FEI varies as a function of followers. In turn, these FEI curves allow us to critically examine the trade-off noted earlier: given a cost structure, what is the optimal level of popularity of an influencer a firm should target to maximize profit from that sponsorship?

We have three key findings. First, FEI, on average across all videos, is positively, but nonlinearly, related to the number of followers. We find that the FEI is, on average, inverted U-shaped in the number of followers: that is, the highest marginal returns on followers tends to be for mid-tier influencers. When thought of in terms of expected impression counts, the average pattern we find is consistent with the S-shaped growth curve postulated in a number of prior studies of advertising (Johansson, 1979). Notably, this causal result differs substantially from a mere correlational, but not causal, analysis, which suggests the biggest marginal effect comes from very small and, more notably, very large influencers. This comparison suggests that, without carefully controlling for video content, heterogeneous effects, and other potential sources of endogeneity, firms may erroneously amplify the importance of mega influencers, and thus make sub-optimal decisions about sponsorships. Our second key finding is that the FEI curves systematically vary with how the firm is trying to engage customers (e.g., entertaining or socializing) and the topics covered in the video (e.g., food or gaming). Thus, it is optimal to seek different influencer popularities for different purposes, which is in line with the industry practice of doing so (albeit typically done in an ad hoc manner based on intuition). Third, we provide counterfactual predictions for how video impressions would grow based on follower size for the sponsored campaigns in our data set. We find there are three types of growth curves – S-shaped, concave, and linear – and showcase why taking an average across campaigns can be misleading for not only the magnitude of FEI but also its relationship with the number of followers. Together, our findings provide firms with a set of practical guidelines for optimizing influencer marketing campaigns, based on sound theoretical underpinnings.

Our work adds to the existing literature in several ways. From a substantive perspective, our work builds on an early but growing literature on optimizing influencer marketing campaigns

(Rajaram and Manchanda, 2020; Yang et al., 2021). In this vein, our work is also connected more broadly to research on electronic word-of-mouth and user-generated content (Kannan et al., 2017). In that area, many studies have examined how the number of connections that individuals have within a network impacts the transfer of information (e.g., Goldenberg et al., 2009; Katona et al., 2011; Liu-Thompkins, 2012; Susarla et al., 2012; Lee et al., 2018). Some have found a positive effect of network size (Goldenberg et al., 2009), whereas others find that a larger number of connections may lead to worse diffusion outcomes (Katona et al., 2011; Liu-Thompkins and Rogerson, 2012). These findings suggest that the effect of network size on diffusion outcomes may be contingent on what is being diffused, motivating us to consider non-linear and heterogeneous effect in our analysis. Other work has recognized the importance of content in social transmission (Zhang et al., 2017), including how content characteristics like emotion (Berger and Milkman, 2012) and topic Hong et al. (2011) affect virality on social media. However, they either assume a linear effect of follower size on reach, or cannot separately identify this effect from other effects of interest, leaving a more nuanced understanding of the relationship between creator popularity and the resulting spread of content largely unexplored.

From a methodological perspective, our work adds to a growing literature on using video data, and unstructured data more generally, in marketing. In particular, video analysis has emerged as a promising tool for understanding the links between the content of video marketing campaigns and firm-relevant outcomes (Li et al., 2019), including in the context of influencer marketing (Yang et al., 2021). In terms of video analysis, our approach aligns with other studies that have considered how various modalities of a video may interact with each other (Rajaram and Manchanda, 2020). However, our work is distinct in how we approach the problem: we leverage a multimodal representation learning framework, which allows us to embed TikTok posts in a lower-dimensional vector space. Our learning framework is based on the variational autoencoder, first introduced by Kingma and Welling (2013), and built on recently in marketing by Dew et al. (2022) and Burnap et al. (2021). Our model is similar in structure to that of the multimodal VAE in Dew et al. (2022), however it differs both in how the multimodal features were processed, and the structure of the model. More specifically, our work leverages modality-specific transfer learning (Zhuang et al., 2020) to process the raw video, audio, and text data, rather than purely hand-crafted features, and a more sophisticated neural network architecture to merge these domains

together.

The second component of our modeling framework, causal machine learning, has also gained significant attention in recent years. Various approaches have been applied in past work in marketing, most notably the causal forest (Guo et al., 2021; Zhang and Luo, 2022). For example, Guo et al. (2021) apply causal forests to assess the heterogeneous treatment effects of physician payment disclosure on the subsequent firms' payments to physicians. In comparison to these methods, Deep IV (Hartford et al., 2017) does not need the "unconfoundedness" assumption, which is less likely to be applicable in our context. It allows for unobserved confounders to affect both the treatment and outcome variables if a valid instrument is present. Deep IV also allows for us to estimate a response curve, to understand how a continuous variables (follower size) affects a continuous outcome (impressions). In that sense, our work echoes earlier work in quantifying advertising response curves in more traditional settings (e.g., Dekimpe and Hanssens, 2007; Zantedeschi et al., 2017).

The rest of the paper is organized as follows. First, we describe our curated TikTok dataset. Then we describe our modeling framework, including our representation learning and causal inference models. After that, we describe several general patterns of results, and the implications of those results, both for managers of influencer marketing, and for our understanding of advertising response more generally. Finally, we conclude with a summary and discussion of some limitations.

2 Data

TikTok provides an ideal empirical context for studying influencer marketing. Besides being one of the most popular social media platforms for firms to sponsor influencers, it also meets the full set of criteria for our causal analysis. More broadly, to provide valid answers for our inquiry around the causal relationship between follower size and video impressions (the treatment effect), the research setting should ideally satisfy several conditions: First and foremost, posts should be collected from a platform where a valid instrumental variable (IV) can be leveraged to infer the treatment effect. Second, social media posts should span a large range of follower sizes and content types, which can identify the heterogeneity in the relationship between the popularity of the

influencer, and the popularity of a video. Third, the growth curve for impressions of posts should be tracked from their introduction (no left censoring) to maturity. Finally, the social media platform should allow brands to collaborate with influencers and the data should contain an identifier for sponsored videos. Our choice of TikTok as our focal platform enables us to meet all of these conditions.

2.1 Data Collection

Specifically, to meet these stringent conditions, we curated a dataset with all TikTok videos, both sponsored and organic, that ever appeared on the public Discover page over a period of six months (from Oct 2020 to April 2021). Our choice of TikTok as the social media platform and the Discover page, in particular, was motivated based on how users are exposed to posts: compared to other browsing channels, the Discover page on TikTok contains videos grouped by categories (termed as hashtags) that are not personalized to any single user's content preferences. The Discover page also updates daily with 1-2 new hashtags, providing us variation in the observed content. Moreover, the Discover page hashtags reflect topics that are currently trending in the TikTok community, and include content creators whose follower size ranges from several hundreds to several millions. Our final dataset consists of 216 hashtags, 30 of them sponsored, and just over half a million videos that were publicly accessible under those hashtags.[4]

For each hashtag that appeared on the Discover page during the data period, we collected key characteristics for every video displayed under that hashtag, including its content, the caption, and any metadata, which, importantly, captures information about the video's creator and the number of views of the video. It is the metadata about a creator that allows us to form our instrumental variable – the average number of hearts (TikTok's equivalent of "likes" on Facebook) received by the creator in the past – which we describe in more detail in the next section. As some videos might contain multiple hashtags, we tracked each video's earliest appearance on the Discover page.[5] Most importantly, we continued to track the number of views for each video on

[4] At first glance, 30 sponsored hashtags may seem like very few. In total during our observation window, there were in fact 58 sponsored hashtags. However, some sponsored campaigns on TikTok have shorter durations (e.g., several days), that made it impossible for us to apply our tracking procedure. The 30 we include are those that we were able to track over at least 14 days, to accurately measure impression counts.

[5] Since we collect all hashtags and all videos associated with them in chronological order, the earliest appearance would capture a video's best possible position to get exposure on the Discover page (i.e., either under the most trending

	Mean	Std. Dev.	Min	Median	Max
Followers	486,169	1,335,649	2	60,102	71,000,679
Impressions	1,038,728	2,468,832	109	4,444,500	165,200,000

Table 1: Summary Statistics
Summary statistics of our key independent variable, follower size ("Followers"), and key dependent variable, impressions ("Impressions")

a daily basis for at least two weeks after its first appearance. This approach allows us to measure, first, how many followers the video's creator had on the day of the video's posting, and how many impressions the video attained subsequently. Moreover, since the Discover page generally features *new* hashtags, it means our panel of observations generally consists of new videos, for which we can track the entire history of the video. Our two week period of observation accounts for more than 90% of a video's one-month views.

TikTok posts are multifaceted and differ from each other in many ways, including the imagery in the video, the background audio, edits and effects added to the video, and the caption under the video. We refer to this set of features, which all fall under the umbrella of unstructured data, as the content features of a post. The other features of a post that we collect we term the post statistics, which include measures like how many hashtags were used and whether those hashtags were trending or not.[6]

2.2 Descriptive Statistics and Exploratory Analyses

In our analysis, the total number of impressions for a video over a period of two weeks operationalizes its reach, our main outcome of interest, and the content creator's follower size when the video is launched is the focal driver of this outcome. Table 1 presents the summary statistics calculated across videos in our data set. The follower size variable shows considerable variation suggesting that our sample contains different types of influencers. The same is the case for the number of impressions suggesting that our data contains a few very popular videos. Due to right skewness for both variables, we use their logarithmic transformation in our analysis. Figure 2 plots their marginal distributions.

Before we introduce our framework for causal inference, it is interesting to see what correla-

hashtag or with the highest ranking if the related hashtags are equally trending).

[6] see the Web Appendix for more details.

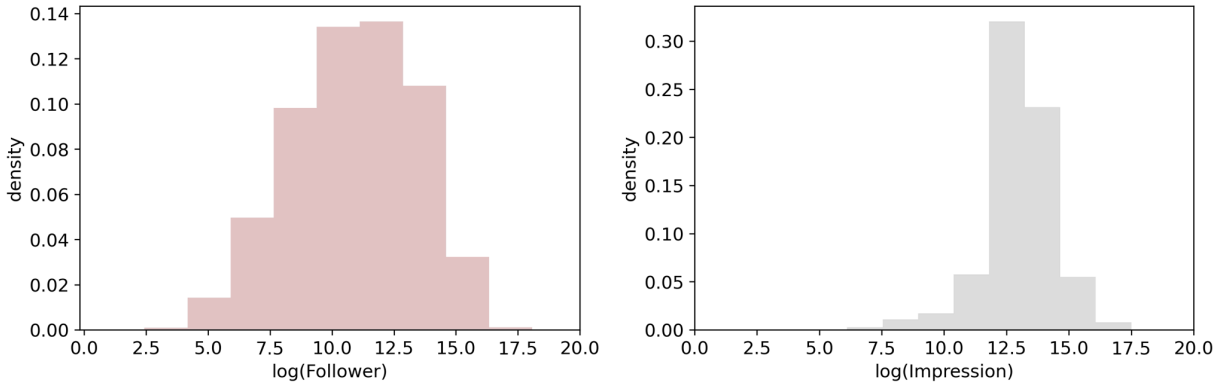


Figure 2: Marginal Distributions
Histograms of the same variables from Table 1, (logged) followers and impressions.

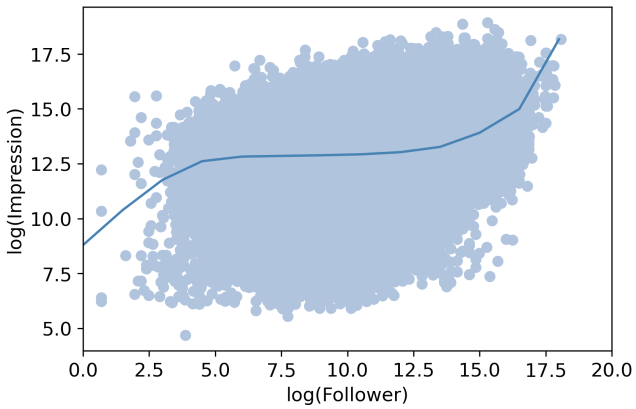


Figure 3: (Correlational) Relationship Between Followers and Impressions
The joint distribution of logged followers and logged video impressions in our data, along with a smoothed estimate of the relationship.

tional patterns exist in the data. Figure 3 shows the relationship between (logged) follower size and (logged) video impressions, in terms of both the raw scatterplot, and a smoothed estimate of the relationship between the two. We see that, correlationally, on average, there is a nonlinear relationship between follower size and the number of impressions. Interestingly, these results suggest that influencers with very small or very large number of followers seem to generate the most significant growth in impressions. A priori, this finding may seem reasonable: mega influencers are extremely popular, and often a target for firm partnerships. However, as we show subsequently, this result is misleading: the simple analysis in Figure 3 does not control for video content or any other sources of endogeneity.

3 Methodology

Our causal analysis between follower size and video impressions on TikTok must address four key issues:

1. **Nonlinear treatment effect:** First, while we expect creators with more followers to have higher reach, this relationship may be nonlinear for a number of reasons. For example, early followers may generally tend to be more "loyal" to a creator, whereas later followers may follow accounts more broadly, leading to higher marginal effects for low follower counts.

2. **The key confound of content:** Second, while we expect a positive *association* between followers and impressions, inferring the *causal effect* of followers requires controlling for the content of a video. To see how content may confound the relationship, consider an example: cat videos may be more likely to receive lots of views, and creators who create cat videos may end up with more followers, because of widespread interest in cats.

3. **Heterogeneity of the relationship:** Third, beyond simply controlling for the content of the video, the actual causal effect of followers may *depend* on the content of the video: given the huge heterogeneity of content on TikTok, it is possible that the reach effect of followers may be stronger for certain content types than others. Therefore, to decide which creators to partner with, firms must first understand the relationship between followers and impressions *for the type of content they want to produce*. Said differently, in the context of the broader causal inference literature, there may be heterogeneous treatment effects across different content types.

4. **Unobserved confounders:** Finally, while controlling for video content is perhaps the most salient challenge, there are a number of other potential confounds that may arise in estimating the relationship between followers and impressions, that are almost certainly unobserved. A subtle but important example is the practice of influencers cross-posting content. Many influencers have accounts across multiple social networks, and it is common practice to use one social network to promote their content on another network.[7] In these scenarios,

[7] https://blog.hubspot.com/marketing/cross-posting

the post on the outside platform may lead to both an increase in followers for the cross-posting creator, and an increase in views for the particular video being advertised. Thus, in inferring the relationship between follower size and video impressions, we need to take into account the potential of unobserved confounders.

To address these four issues, we propose a framework based on two components: first, to summarize the content of the video, and thereby partly address points (2) and (3), we propose a representation learning model for TikTok posts. This model takes as its inputs TikTok posts, and outputs a dense vector representation of the post's content, which, for the remainder of this section, we will refer to as $\boldsymbol{r}$. Second, conditional on $\boldsymbol{r}$, and any other observables about the creator and post, we leverage a machine learning-based causal inference framework, together with an instrumental variable informed by how users find content on TikTok, to model the relationship between followers and impressions. Based on the Deep IV framework proposed by Hartford et al. (2017), our model allows us to simultaneously address all four key issues.

3.1 Learning Representations of TikTok Content

We need to distill the essence of a TikTok post into a vector of features that can be seamlessly integrated in our causal model. These features should capture both what the post is about, and the overall quality of its content. We learn such a feature vector by leveraging a combination of two technologies: transfer learning based on pre-trained feature extractors, and multi-modal representation learning. With transfer learning, we extract a set of features that describe all aspects of a TikTok post, including its video and accompanying text. The extraction of features using pre-trained models simplifies the subsequent representation learning process, and allows us to explicitly incorporate especially relevant features for TikTok posts (e.g., how the posts are edited). Then, our multi-modal representation learning framework condenses this set of features into a single, dense, lower-dimensional vector, which we call $\boldsymbol{r}$, that captures the links across them, and can be used in a causal model. We briefly describe the two parts, and defer a detailed discussion to the Web appendix.

Extracting Features with Transfer Learning From each post on TikTok, we extract features from four modalities: (1) textual features, from both the video and the caption attached below; (2)

image features, which are learned from the image frames of the video; (3) audio features, which capture features of the video's sound; and (4) editing features, which capture how creators edit their videos on the platform. While the first three are standard in video analysis, the fourth is a modality unique to TikTok. The specific features we extract include:

- **Text:** Our textual features are based on three sources: the video's caption, any text overlaid on the videos themselves (referred to as *stickers*), and text capturing the words spoken in any video voice-overs. Across all three text modalities, we further process the raw text into word embeddings, leveraging the pre-trained Word2Vec model (Mikolov et al., 2013), yielding our final set of textual features.

- **Image:** The visual component of video data consists of sequences of image frames. In our analysis, rather than extracting and analyzing each individual image frame, of which there may be thousands, we extract image frames in five second intervals. By doing so, we greatly reduce the volume of data, while still maintaining a relatively rich set of image frames. To that collection of frames, we apply two pre-trained neural networks: VGG-19 (Simonyan and Zisserman, 2014), which is an object recognition model, and SentiBank (Borth et al., 2013), which predicts the emotional qualities of images. From VGG-19, we first extract the second-to-last layer of the model for each frame of the video, which has been shown in prior work to be a good, general-purpose representation of image content (e.g., Wei et al., 2019; Zha et al., 2015). These features are then synthesized using an LSTM layer in our model, which accounts for their sequential structure. We also compute an average of the final layer of VGG-19, averaged across frames, which predicts which of 1,000 potential classes of objects may be present throughout the video. Finally, from SentiBank, we extract the final model prediction over 2,089 emotion classes, again averaged across frames, giving us an average sense of the emotional content of a video.

- **Audio:** While audio is an important feature of all video data, it is especially important on TikTok, as many videos are posted with a common audio track, where the audio motivates the video content. We extract a number of audio features from each video including the universality of the track, meaning how common it is across TikTok videos, different sound classes based on the YAMNet model (Gemmeke et al., 2017; Hershey et al., 2017), and other

high-level acoustic features, capturing the raw traits of the sound itself (e.g., intensity and pitch).

- **Editing:** TikTok offers various tools that allow users to easily add special effects to their videos, including simple edits such as video length and speed, and more complex edits such as image filters. To measure how a video has been edited, we develop a number of custom feature extractors, that capture the video length, whether the video has overlaid images (called "stickers"), the average scene length, the variance of features across frames, and whether there are any filters applied to the video.

In sum, our feature extractors yield four sets of features, that capture the content of a TikTok post. More details are available about these features in the Web Appendix.

Representation Learning While many of the features we extract are already relatively high-level, the full set of features is still very high-dimensional, and each modality is treated separately by the feature extraction procedure. To further reduce the dimensionality of these features, and synthesize them in a way that produces a high-level representation of the essence of each post as a whole, we apply a representation learning procedure based on the variational autoencoder.

Variational autoencoders (or VAEs; Kingma and Welling 2013, Rezende et al. 2014) are based on traditional autoencoders, which are machine learning models with two components: an encoder that compresses the data to a dense vector representation, and a decoder that reconstructs the original data from that representation. The variational autoencoder is a probabilistic variant of this framework, where the generative process of the observed data, $\mathbf{x}_i$ for observation i, is modeled as a function of latent, lower-dimensional vector representations, $\boldsymbol{r}_i$. Mirroring the classic autoencoder, the VAE has two parts: the encoder (or "inference network") specifies an approximate posterior distribution over $\boldsymbol{r}$, given data $\mathbf{x}$, by learning a density $q_\phi(\boldsymbol{r}|\boldsymbol{x}) \approx p_\theta(\boldsymbol{r}|\boldsymbol{x})$. The decoder (or "generative model") models the data generating process as a function of $\boldsymbol{r}$, $\mathbf{x} \sim p_\theta(\boldsymbol{x}|\boldsymbol{r})$. From the decoder, we can see that the $\boldsymbol{r}$ acts as a sufficient statistic for the data: given $\boldsymbol{r}$, we can reconstruct $\boldsymbol{x}$. In both components, the link between the data and the representation is parameterized by neural networks, which have parameters ϕ for the encoder, and θ for the decoder. VAEs have been shown to outperform traditional (non-probabilistic) autoencoders in learning meaningful,

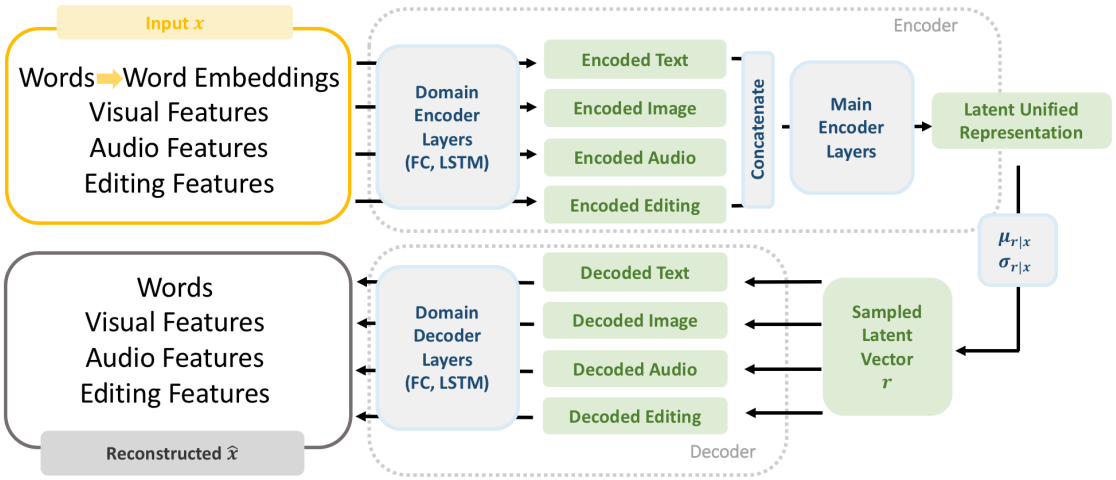


Figure 4: Network Architecture of the Proposed VS-VAE
Illustration of the network architecture of our representation learning framework, the VS-VAE. At top is the encoder, and at bottom, the decoder.

low dimensional representations of data, across a variety of domains (Hsu et al., 2017; Yao et al., 2019).

The version of VAE we develop is a *multimodal* VAE. Its structure is similar to a standard VAE, with both encoders and decoders, but it incorporates special structure within each of these components that allow it to merge together information from each of the four modalities of TikTok data (text, image, audio, editing). Specifically, in the encoder, we first use modality-specific networks that capture the specifics of a given modality. Next, we combine the modality-specific encoders using a neural network, which allows them to interact in a potentially nonlinear manner. This joint enconder is then used to estimate the latent representation of a post. The decoder mirrors this structure. Given the highly structured nature of the modality-specific encoders and decoders, we refer to our proposed model as a "very structured" VAE, or VS-VAE. Figure 4 shows the overall modeling framework. The primary output of this framework comes from the encoder: given a new TikTok post $\mathbf{x}_i$, the encoder gives us an estimate of $p(\boldsymbol{r}_i|\mathbf{x}_i)$ in terms of a posterior mean, μ_i, and variance, σ_i. This posterior mean is the representation we use to control for post content. Going forward, we will slightly abuse our notation and use $\boldsymbol{r}$ to refer to this mean. We include a detailed description of each component of our VAE framework in the Web Appendix.

Validating the Representations We validate that our representations capture the content and quality of TikTok posts in three ways. First, if the VS-VAE is indeed capturing post content, then

	Text Accuracy	(Avg.) Image MSE	Audio MSE	Editing MSE
VS-VAE	0.38	10.81	0.97	0.69
NIR	0.04	35.21	3.96	40.33

Table 2: Reconstruction Statistics under VS-VAE and NIR
For text, we measure the percentage of the reconstructed words that correctly match the original words (a higher number indicates better fit). For the other three modalities, we calculate the mean squared error between the reconstructed and original features (higher numbers indicate worse fit). For image, we average the fit statistic over all image frames.

for a new post, we should be able to use the VS-VAE to accurately reconstruct the post from its representation. We test this out-of-sample reconstruction, and find that the model is indeed able to accurately reconstruct what it is given. We report the results in Table 2, where we compare the reconstruction ability of our model to a no information rate (NIR). For the NIR, we simply predict the mean value for every feature. The comparison to the NIR gives us a way of assessing the magnitude of the reconstruction statistics. In short, in comparison to the NIR, we see our model is able to accurately reconstruct the data, giving us confidence that the representations do indeed capture the content of videos.

To further validate that the representations meaningfully capture video content, we examine the distribution of videos in $\boldsymbol{r}$-space. While the dimensions of $\boldsymbol{r}$ are not interpretable, distances between videos in $\boldsymbol{r}$ are: posts that are closer in terms of $\boldsymbol{r}$ are expected to have similar features. Thus, if $\boldsymbol{r}$ meaningfully captures the content of posts, we should find that the closest videos to a focal video in $\boldsymbol{r}$-space are similar to that focal video. In Figure 5, we show the top three nearest neighbors in the latent $\boldsymbol{r}$ space for a set of three focal videos under three randomly selected hashtags. For each video, we show five frames, as well as the hashtag that the video was posted under. Overall, we see there is a high degree of similarity among neighboring videos. For example, the focal and neighboring videos under #FoodTikTok all start with a dish, followed by the cooking process, and then show the dish again. The color palettes across the frames are similar as well. Indeed, across all focal videos, we see that, visually, the neighboring videos follow a similar trajectory, color tones, and objects presented throughout. Less obvious from Figure 5, we also see that neighboring videos share many other features, including the music used, the extent to which stickers, or overlaid textual comments, are used, and the scene length.[8]

The first two analyses illustrate that the representations can capture video content, but it is not

[8]We again defer more details to the Web Appendix.

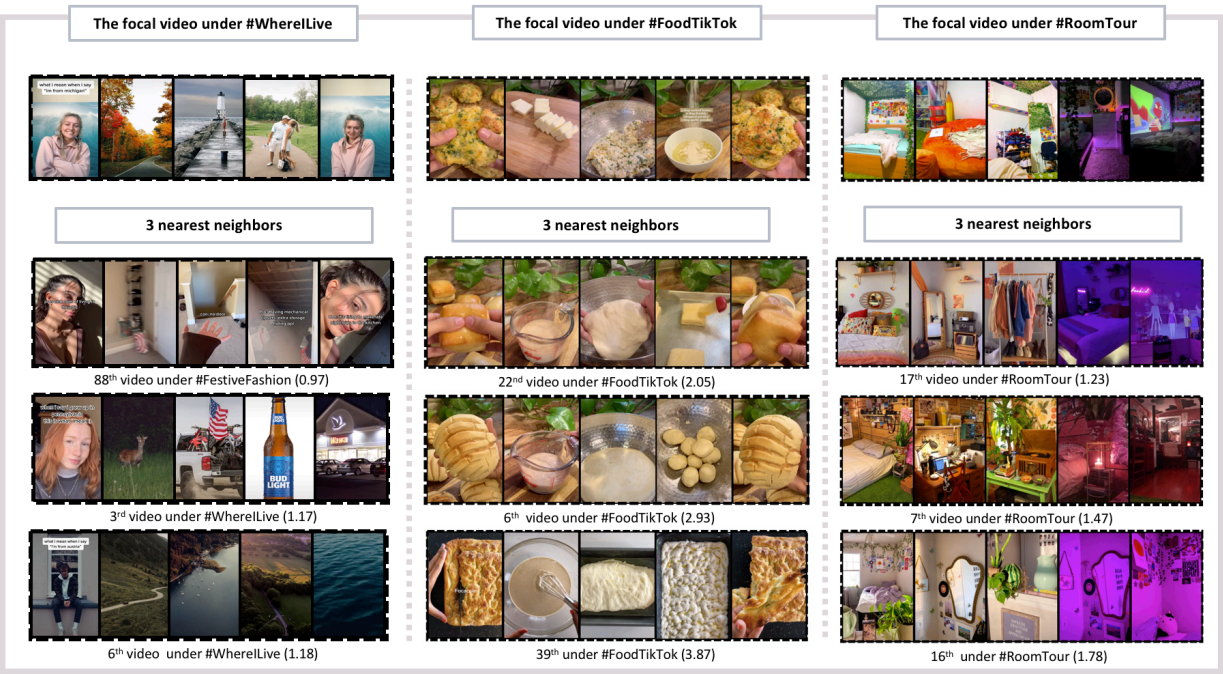


Figure 5: Nearest Neighbors Case Studies
The three nearest neighbors in $\boldsymbol{r}$-space for three randomly selected focal videos, selected from different hashtags.

clear that they capture the underlying *quality* of a video. While it is difficult to measure quality, one metric we can use is video popularity: can we predict which videos become surprisingly popular (or surprisingly unpopular) using $\boldsymbol{r}$? By surprisingly, we mean that a video outperforms our expectations, given the hashtag the video is posted in, the popularity of its creator, and its other initial conditions, in terms of its impression growth. To do this, we first calibrate a hierarchical logarithmic growth model for impressions over time, with post-level parameters capturing the initial number of impressions for the post, and its growth rate over time. We include the initial conditions of the video as predictors of these post-level parameters. We then test whether the representations $\boldsymbol{r}$ can incrementally explain the post's initial impressions and growth rate, beyond just the initial conditions. Indeed, we find they can: our representations can predict with over 70% accuracy which videos by over- and under-perform their initial conditions. We include complete details of the impression growth model model and our testing procedure in the Web Appendix.

3.2 Deep Instrumental Variables

While the representation learning procedure provides a way to control for video content, there are three remaining concerns for our causal analysis: a potentially nonlinear relationship between followers and impressions; potentially heterogeneous response functions across content; and the need to account for possible unobserved confounders. To account for these concerns, we turn to deep instrumental variables.

Instrumental variables methods have been widely used for causal effect estimation with observational data, when there may be unobserved confounders. Most existing approaches (e.g., two-stage least squares) do so under two strong assumptions: linearity and homogeneity (e.g. Angrist et al., 1996). That is, the treatment affects all individuals in the same constant way. In our case, as described above, both of these assumptions are unlikely to be true. To overcome these limitations, a recent literature has emerged blending together standard causal inference approaches with flexible machine learning models (e.g. Chernozhukov et al., 2017; Athey et al., 2019; Farrell et al., 2021). In the context of instrumental variables, Hartford et al. (2017) have proposed Deep IV, a model that generalizes two-stage least squares by estimating both the first and second stages of the IV framework through deep neural networks, thereby allowing both heterogeneous and nonlinear estimation of causal effects. As Deep IV is, to our knowledge, new to marketing research, we will now summarize the Deep IV framework, and connect it to our setting, following closely the exposition in Hartford et al. (2017).

Deep IV assumes the following general model for the data generating process:

$$y = g(t, x) + e, \tag{1}$$

where y is the outcome variable; t is the treatment variable; x is a set of observed features that affect both t and y; z is the instrument; e are some unobserved variables, i.e., the error, that might be correlated with x, t, and y; and $g(\cdot)$ is a flexible function of both x and t that captures the relationship of interest. This model essentially assumes a potentially nonlinear relationship between the treatment, controls, and the dependent variable, while assuming that any unobserved effects enter in additively. In our context, y is the number of impressions a video receives, t is the number of followers the post's creator has, and x contains both the content features of the video ($\boldsymbol{r}$) and

any other observables about the post and its creator that are not included in $\boldsymbol{r}$.

Our goal is to predict counterfactual values of y under different treatment values t, conditional on observables x. That is, in our context, we want to understand the reach of a video, if the same video were posted by different creators who vary in their follower sizes. To address such a question, Hartford et al. (2017) define a counterfactual prediction function,

$$h(t, x) \equiv g(t, x) + \mathbb{E}[e \mid x], \tag{2}$$

where, crucially, the expectation of e does not depend on t. Given two treatment values (i.e., two follower sizes), $h(t_1, x) - h(t_2, x) = g(t_1, x) - g(t_2, x)$, suggesting that, if we are able to estimate $h(t, x)$, we can infer these counterfactual quantities. However, estimating $h(t, x)$ is not straightforward. We cannot, for instance, simply try to fit the model to the data: under Equation 1,

$$\mathbb{E}[y \mid t, x] = g(t, x) + \mathbb{E}[e \mid t, x] \neq h(t, x),$$

as we have allowed $\mathbb{E}[e \mid t, x] \neq \mathbb{E}[e \mid x]$.

To get an unbiased estimate of $h(\cdot)$, Hartford et al. (2017) suggest using an instrument. Specifically, as in the standard instrumental variables setting, suppose there exists an variable z, such that z satisfies the following three conditions:

1. **Relevance:** $F(t \mid x, z)$, the distribution of t given x and z, is not constant in z;
2. **Exclusion:** z does not enter Equation 1;
3. **Exogeneity:** z is conditionally independent of the error, e;

then we can leverage this instrumental variable to estimate $h(\cdot)$. Specifically, given z, we can write:

$$\mathbb{E}[y \mid x, z] = \mathbb{E}[g(t, x) \mid x, z] + \mathbb{E}[e \mid x] = \int h(t, x) dF(t \mid x, z), \tag{3}$$

which allows us, theoretically, to estimate $h(t, x)$ given two observable quantities: $\mathbb{E}[y \mid x, z]$ and $F(t \mid x, z)$.

Until now, the derivation has closely followed the standard derivation of instrumental variables approaches (e.g., Newey and Powell, 2003). The Deep IV framework departs from standard

approaches by operationalizing Equation 3 directly. Hartford et al. (2017) propose directly optimizing an estimator $\hat{h}(\cdot)$ for $h(\cdot)$ by minimizing the objective function,

$$\min_{\hat{h}\in\mathcal{H}} \sum_{n=1}^{N} \left(y_n - \int \hat{h}(t, x_n) dF(t|x_n, z_n) \right)^2, \tag{4}$$

where $\mathcal{H}$ is defined as a set of functions that can be parameterized by a neural network, allowing $\hat{h}$ to be rewritten as $\hat{h} = h_\zeta(t, x)$, where ζ are the parameters of the neural network. Unfortunately, without knowing $F(t \mid x, z)$, this equation cannot be used directly. Hence, Hartford et al. (2017) propose first estimating $F(\cdot)$ with a parametric distribution function $\hat{F}(t \mid x, z)$. To allow maximal flexibility in the form of $\hat{F}$, Hartford et al. (2017) again parameterize it using neural networks, specifically assuming, for the case of a continuous t, that $\hat{F}(t \mid x, z)$ is the distribution function of a mixture of Gaussians, whose parameters depend on the data through a neural network with parameters η, subsequently denoted as $\hat{F} = F_\eta(t \mid x, z)$. Thus, the Deep IV approach replaces the usual first stage of the two-stage least squares procedure by computing a flexible estimate $F_\eta(t \mid x, z)$ of the distribution of the treatment given observables. Then, conditional on the learned η from stage one, denoted η^*, Deep IV solves

$$\zeta^* = \arg\min_{\zeta} \frac{1}{N} \sum_{n=1}^{N} \left(y_n - \int h_\zeta(t, x_n) dF_{\eta^*}(t|x_n, z_n) \right)^2 \tag{5}$$

to learn an estimate of $\hat{h}(\cdot) = h_{\zeta^*}(\cdot)$, which can be used to quantify counterfactuals of interest. In practice, both of these models (for h and F) are trained using stochastic gradient descent, and we refer readers to Hartford et al. (2017) for additional implementation details.

Applying Deep IV to TikTok Data We now connect the general Deep IV framework to our particular empirical setting. To estimate the causal impact of an influencer's popularity, operationalized as their follower size, on the reach of a post, operationalized as video impressions, we assume the following data generating model:

$$\log(R_j) = g(\log(F_{i(j)}), x_j) + e_j, \tag{6}$$

where R_j is the reach of video j, defined as how many impressions the video had at the end of our tracking period; $F_{i(j)}$ is the follower size of creator i at the time of posting video j; $i(j)$ is the one-to-one mapping between video j and the influencer who posts video j; x_j is a vector that includes both summary statistics of this post and its video (e.g., how many hashtags video j uses; see the Web Appendix for more details) and the representation, $\boldsymbol{r}$, described previously; and e_j is an error term. We model reach and followers on the log scale to account for their heavy skewness (see Figure 2, and to enable easy calculation of elasticities, as we describe in a later subsection. To estimate counterfactual quantities under this model, we need an instrument for followers. We describe the instrument we use — past video success — in detail in the following subsection. Given that instrument, the rest of our implementation of Deep IV largely follows Hartford et al. (2017). We give more details, including our assumed model and network architectures for h and F, in the Web Appendix.

3.3 Instrument: Past Video Success

All of the preceding exposition assumes that we have a valid instrument for followers on impressions, with which to estimate the counterfactual prediction equation with Deep IV. Such an instrument should cause variation in an influencer's follower size (the "relevance" condition), but not directly affect the impression count of the current (focal) video (the "exclusiveness" condition), nor be correlated with the error term (the "exogeneity" condition). We propose that a creator's *past* video performance meets those criteria. Specifically, given that we do not have a single official measure of video performance, we use a video's hearts count, which is TikTok's equivalent of Facebook's "Likes," as a practical proxy. Consequently, we define an influencer's past video performance as the average number of hearts received among all their videos posted up to the time of the current post.

Relevance We first aim to demonstrate that an influencer's past video performance indeed affects their follower size. Intuitively, high past video performance signals that an influencer can create high-quality content, and thus tends to attract more people to follow them. However, we can also quantitatively establish relevance: for simplicity, let Past denote the IV, and let Follower denote the number of followers a creator has at the time of posting. We find that the correlation

between Past and Follower is fairly high, 0.736. Moreover, regressing Follower on Past, we obtain a statistically significant coefficient for Past at 0.001, and a large R-square (0.542), further illustrating that a substantial proportion of variation in an influencer's follower size can be explained by their past video performance. Thus, the instrument is relevant.

Exclusiveness and Exogeneity Having established the relevancy of the proposed IV, we now argue that the instrument also meets the more difficult conditions of exclusiveness and exogeneity. Intuitively, exclusiveness (or the exclusion restriction) means the instrument only drives the outcome through the treatment, and exogeneity means the instrument is uncorrelated with the error term (i.e., any potential confounders). In our context, to establish these conditions, we fist logically argue that past video performance only affects current video performance through the number of followers of the creator. To establish that, we first need to understand how users discover content on TikTok. On TikTok, there are five possible channels through which users can discover content: (1) TikTok's personalized "For You" feed; (2) the Discover page,[9] (3) the Following feed, (4) searching on TikTok, and (5) sharing. We now explore each of these channels, and show how past video performance does not drive discovery through any of them, except through the mechanism of following.

1. **For You:** The first channel, the For You feed, is a customized list of videos recommended by TikTok based on each user's activities and interests on the platform. According to TikTok, whether or not an account has had previous high-performing videos is not a direct factor in the recommendation system.[10] Therefore, from the perspective of *direct* content recommendation, an influencer's past video performance will not affect how TikTok promotes his/her current video and thus its impression. However, there may still be *indirect* targeting: high past video performance might be the result of an influencer posting content of a particular kind that is popular or trendy (e.g. cooking, dancing). TikTok then might be more likely to target users with a current video of the same type, based on popularity of the content. However, since we control for the video content through $\boldsymbol{r}$, this mechanism will not violate our identification.

[9] https://newsroom.tiktok.com/en-us/discover
[10] https://newsroom.tiktok.com/en-us/how-tiktok-recommends-videos-for-you

2. **Discover Page:** TikTok's Discover page shows users a list of hashtags (i.e., video categories) that are currently trending. After clicking a hashtag, users are taken to a ranked listing of videos under that hashtag. Both the list of hashtags and the videos under each hashtag are the same for every user. In terms of our instrument, the primary concern is that past video performance affects exposure of a video on this page. This could happen in one of three ways: (1) through the hashtag (part of the video caption) itself (i.e., past success under a hashtag causes an influencer to post again under a related tag), (2) through the ranking of the video under a hashtag (i.e., creators with more successful past videos are ranked more highly), or (3) through the thumbnail of the video visible on the discover page (which may be informed by a creator's past success). We control for all of these in our model: (2) is included directly as covariates as part of x in Equation 6, and both (1) and (3) are accounted for through the video representation $\boldsymbol{r}$, which, among other things, captures the visual content of the video.

3. **Following Feed:** The following feed is, perhaps, the easiest to address: when a user follows a creator, that creator's videos will appear on the user's feed. It is not possible to appear on the following feed without following; thus, exposure to videos here happens exclusively through the follower count.

4. **Searching:** If a user is interested in some particular content or influencer on TikTok, they can search them directly using TikTok's search bar. If the search relates to the content of the video, this will be captured directly in our model through the post features, $\boldsymbol{r}$. If, on the other hand, the user searches for the influencer directly, this suggests the user already knows of the existence of the influencer, which likely happens either when the influencer is famous, or when there are some concurrent exogeneous shocks that suddenly make the influencer to widely known. In the first case, prior fame will be captured by the influencer's follower count; in the second case, exogeneous shocks are not problematic, insofar as they are, by definition, unrelated to past video performance.

5. **Sharing:** Other than these four direct channels, a video might also get impressions indirectly by being shared to others. The concern here is, again, if past video success directly drives impressions via sharing. Given users may share similar content over time, and influencers

may tend to post similar content over time, influencers who have historically been more successful might continue getting more shares for their current videos. This concern is alleviated in our context, since the driving effect is the similarity of the content, and we control directly for content. Alternatively, regardless of content, people may have a higher propensity to share because of the influencers themselves, especially if the influencer is a celebrity. To control for celebrity status, we include whether an influencer is verified or not as part of our post statistics (i.e., part of x).[11].

In summary, past video performance, operationalized by number of hearts given, on average, to a creator's past videos, is a logically valid instrument for measuring the impact of number of followers on reach: given how people find content on TikTok, past video performance only affects the number of impressions of current videos through the number of followers a creator has, or through other factors that are explicitly controlled for, making it conditionally independent of the error term in our model.

Validity Test Finally, to further empirically test the quality of past video success as our IV, we perform a validity check recently proposed in the machine learning literature by Li et al. (2022). The proposed method leverages machine learning to validate the typically impossible-to-check exclusiveness and exogeneity assumptions in a nonlinear IV framework. Generally speaking, if the instrument can predict the error term of the focal model, then it cannot possibly be a valid instrument. Therefore, we check whether the error term from Equation 1, which should be independent from the instrument, can be predicted by the instrument z and known covariates x with a machine learning model any better than a model that is identically zero. Let $\hat{e}$ denote the estimated error term in our Deep IV framework. We then separately estimate $\hat{e}$ using another machine learning model, using x and z as inputs. We then compare the MSE loss of predicting $\hat{e}$ using x and z, with the MSE loss from a model that predicts identically zero. Specifically, denoting the former as ℓ and the latter as ℓ_0, we compute

$$\gamma = 1 - \frac{\ell}{\ell_0}. \tag{7}$$

[11] TikTok has a full list of criteria to determine whether an influencer could be rewarded as verified which contain a large variety of aspects that might make an influencer special, including being a celebrity (https://newsroom.tiktok.com/en-us/how-to-tell-if-an-account-is-verified-on-tiktok)

If γ is close to zero, it means the loss of the model predicting $\hat{e}$ with x and z is close to a model that just predicts 0, which is exactly what we expect if the instrument is valid. Indeed, in our case, we find $\gamma = 0.00023$, offering empirical evidence on the validity of our instrument.

3.4 Deriving Follower Elasticities from Deep IV

Given this instrumental variable, the output of our Deep IV framework is an estimate h_{ζ^*} of the counterfactual prediction function, as defined in Equation 2. Intuitively, the counterfactual prediction function lets us understand how (log) impressions of a video causally vary as a function of the popularity of the influencer that posts the video. We now show how the counterfactual prediction function can be used to derive our key quantity of interest: the follower elasticity of impressions.

More specifically, in our context, h_{ζ^*} is a function of both log followers, and the set of post characteristics x (which contains the post representation $\boldsymbol{r}$): $h_{\zeta^*}(\log(F), x)$. As described in Equation 2, this counterfactual prediction function is *not* the function $g(\cdot)$ from the data generating process, which is the true object of interest, but the sum of $g(\cdot)$ and the error term conditional on x. However, we can still use h to predict the reach of a post defined by x under different values of $\log(F)$, by taking differences of h. That is, concretely, suppose we have a video with features x, and we want to understand how many more impressions we would get on the video if it were posted by an account with $\log(F')$ followers, as opposed to $\log(F)$ followers. To do that, we can simply compute:

$$h_{\zeta^*}(\log(F'), x) - h_{\zeta^*}(\log(F), x).$$

Because the underlying model for h is a neural network, the learned h captures both nonlinear effects between F and R, and interactions between F and x in determining R. In this way, h allows us to understand how the effect of F on R may vary by the covariates, x, or what the broader literature has referred to has heterogeneous treatment effects (e.g., Imai and Ratkovic, 2013; Wager and Athey, 2018).

In our context, the "treatment" is continuous, rather than the more common setting where the treatment is a discrete intervention. Thus, when describing the relationship between F and R given x, we need to derive functional quantities, that capture the *marginal* effect of increasing

F on R. To do that, we again consider taking differences of h. Suppose we are considering how the reach of a video with features x would change if we increased its log followers from $\log(F)$ to $\log(F) + \epsilon$, $\epsilon > 0$. By the definition of h, we can compute the expected change in the reach of the video resulting from that ϵ increase by:

$$\mathbb{E}[\log(R) \mid \log(F) + \epsilon, x] - \mathbb{E}[\log(R) \mid \log(F), x] = h_{\zeta^*}(\log(F) + \epsilon, x) - h_{\zeta^*}(\log(F), x). \tag{8}$$

In the limit, as $\epsilon \to 0$, we can use this to derive marginal effects of the number of followers on the reach of a post, which we refer to as the *heterogeneous marginal effect*, or HME. Mathematically, the HME corresponds to:

$$\text{HME}(F \mid x) = \frac{\partial}{\partial \log(F)} \mathbb{E}[\log(R) \mid F, x]. \tag{9}$$

In practice, we compute this numerically, by computing the change in the estimated h_{ζ^*} for small values of ϵ. We call this the heterogeneous marginal effect because it depends on x, the post features. That is, the marginal effect of increasing followers may be different, depending on what type of video we are posting.

Finally, based on these marginal quantities, we can compute what we term the *follower elasticity of impressions*, or FEI, which is a central quantity of interest in determining optimal influencer selection, as we will demonstrate later. Specifically, the follower elasticity of impressions is the percentage change in impressions corresponding to a percentage change in the video's creator's followers (analogous to the common price elasticity). Mathematically, given x, in the neighborhood around a given follower value F_0, we can approximate the relationship between R and F by:

$$\log R \approx \text{HME}(F_0 \mid x) \log F + c(F_0), \tag{10}$$

where $c(F_0)$ is a term that does not depend on F. In words, we see that, pointwise, the HME is the coefficient of a log-log model of R on F. From this, we can see that the HME is, in fact, equivalent to the follower elasticity of impressions, conditional on video features x. That is, we can interpret the HME as the percentage increase in impressions associated with a unit percentage increase in followers. Since the HME varies at different levels of followers, the FEI will also vary. Hence, when examining the FEI, we often plot what we call the FEI curve, which is simply the value of

the FEI at different values of followers. As we will show in the next section, these FEI curves, and more generally this interpretation of the HME and an elasticity, is crucial in determining whether a firm will benefit from sponsoring an influencer with a higher number of followers.

3.5 Benefits of Our Framework

Before describing our results, we first briefly emphasize the key benefits of our combination of representation learning and Deep IV. The key to our framework is that the mechanisms behind content discovery on TikTok yield a plausible instrumental variable. Given the potential for unobserved confounders, and the lack of experimental variation in the data, utilizing this IV is essential for computing our counterfactual predictions. While simpler IV methods, including the classic two-stage least squares approach, could have been utilized, such methods would not yield nonlinear or heterogeneous estimates of the (continuous) treatment effect of popularity on reach, which are essential in our context, and will be the focus of many of our results in the following section.

Deep IV has many benefits, but it is not the only nonparametric IV estimator proposed in the literature. Compared to other nonparametric IV methods like sieve- or kernel-based methods (Newey and Powell, 2003), Deep IV does not require a strong prior understanding of the DGP (data generating process) by the researcher and stays computationally tractable, especially when there are more than a handful of inputs and a large number of training samples, by employing deep learning algorithms. More recent work by Farrell et al. (2021) has further established the convergence properties of deep learning-based frameworks under various network architectures and a general class of nonparametric regression-type loss functions.

Finally, in our framework, we use representation learning to capture the content of videos before doing our causal estimation. In contrast, one could, in theory, avoid this stage by utilizing the video data directly as the covariates x in the Deep IV framework. We favor the representation learning approach primarily because optimizing the Deep IV model directly with the raw, large-dimensional video inputs is infeasible. Despite trying numerous architectures, when using the raw video data in the Deep IV estimation, we were never able to achieve satisfactory convergence. In this sense, our representation learning approach plays a role similar to unsupervised

pre-training, which prior work in the deep learning space has found can help avoid overfitting and convergence issues (Erhan et al., 2010). A second benefit of the representation learning approach is that the dense vector representations of videos can be useful for validating the meaningfulness of the inputs, and for performing other validation analyses, as we illustrated in an earlier subsection.

4 Results

We begin by discussing our results—our heterogeneous follower elasticity curves—at a high level, before describing their implications for designing successful influencer marketing campaigns.

4.1 Follower Elasticity Curves

To begin, in the left panel of Figure 6, we plot the average FEI curve for all videos in our dataset, averaging over all video types (i.e., over all x). We see that, overall, the FEI is positive, suggesting that additional followers can always generate more impressions, though at different rates. We find the average FEI follows an inverted U-shape, suggesting that a percentage gain in followers yields, on average, the largest percentage gain in impressions for relatively small influencers: the maximal average FEI is attained at just 4,000 followers. To understand what this inverted U-shape implies about the overall reach of a hypothetical video, in the right panel of Figure 6 we convert the FEI curve into an expected *counterfactual prediction curve*. This curve tells us how many impressions we expect a video with a given number of followers to achieve.[12] We see that the counterfactual prediction curve is S-shaped. This result is very different from the pattern implied by the descriptive analysis, in Figure 3, which suggested that the highest marginal rates of return for followers are in the low and high ranges.

While the average curve is suggestive, the real power of our framework comes from its ability to compute FEI curves for any value of x, allowing us to see how the FEI varies as a function of what's actually being posted. We first consider different engagement tactics firms may use in their campaigns on TikTok. In some cases, firms use challenges to educate consumers on how best to

[12] As described in the previous section, the FEI is essentially the derivative of the counterfactual prediction curve. Hence, to convert an FEI curve to a counterfactual prediction curve requires us to make some assumptions about the overall level of impressions of the video (i.e., the constant from integration). For average curves, like in Figures 6, we assume the constant is just the average number of impressions of all videos in the data. For later cases, we use the actual level of impressions observed for that video.

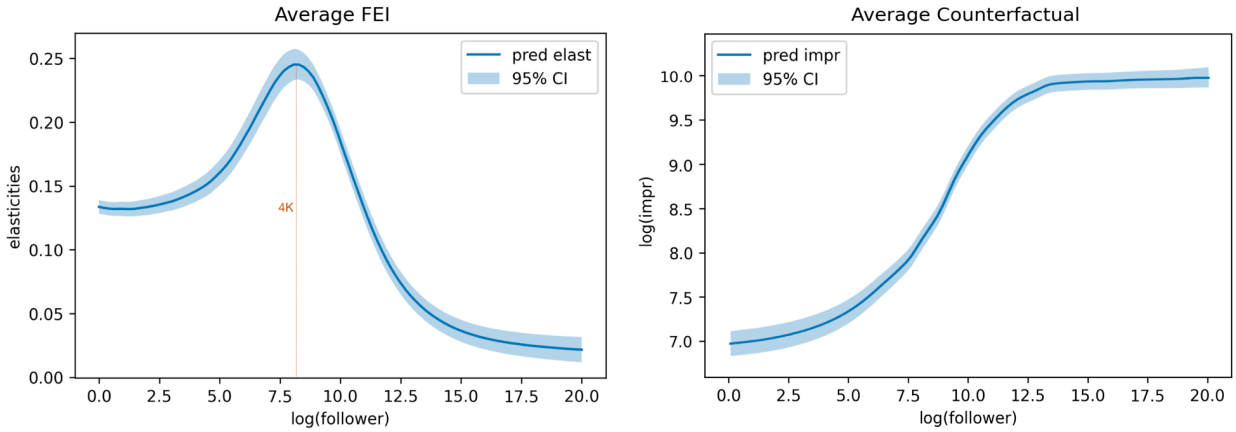


Figure 6: Average FEI and Counterfactual Prediction
At left, we plot the average follower elasticity of impressions, averaged over x, computed at different levels of log followers. At right, we plot the counterfactual prediction curve, which is the average predicted impressions of a video, given its number of followers. Uncertainty bands are 95% confidence intervals for the mean, computed by bootstrapping.

use their products, while at other times, they may integrate their brands into social activities that encourage people to participate. These different challenge types mirror a classification scheme previously proposed by Dolan et al. (2019) for how firms can engage customers on social media. In particular, Dolan et al. (2019) classify social media posts into the categories informational, entertaining, and socializing. Table 3 gives specific definitions for these labels, and several examples hashtags matching each. Using the heterogeneous FEI curves enabled by our framework, we can explore whether different types of influencers may be more effective for different engagement tactics. To that end, we first recruited independent raters to assign the informational, entertaining, and socializing labels to each of our 216 hashtags. Then, we computed an average FEI curve for all videos from the hashtags belonging to each of those categories. Figure 7 shows the results: we find that there are significant differences in FEI based on the engagement tactic. Influencers with less than 10K followers achieve faster impression growth for entertaining and informational hashtags, while the influencers with a larger number of followers do so for socializing hashtags. The result has face validity as the spread of social content relies more so on the follower size of influencers.

Engagement Tactic	Definition	Examples
Informational	Content that provides users with resources and helpful information (e.g., content about events, places, opportunities, people or celebrities).	#WomenInStem, #HomeOffice, #FallDIY
Entertaining	Content that is meant to be fun and entertaining, without explicit informational value	#GreenScreenScan, #FallGuysMoments, #OhNo
Socializing	Content that encourages users to interact with one another, and stimulates their desire for social integration and social benefits	#LaughingDuet #PerfectMatch #GroupChat

Table 3: Engagement Tactics
Definitions and examples of the three engagement tactics we use to classify hashtags.

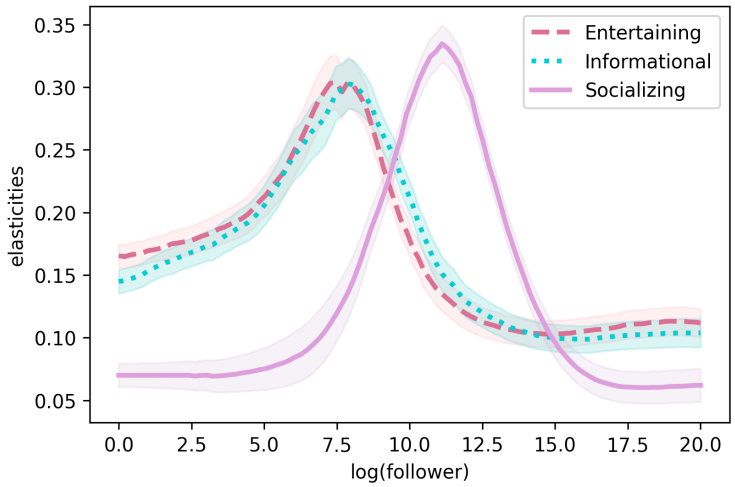


Figure 7: Average FEI by Engagement Tactic
The average FEI for videos that were labelled as coming from entertaining, informational, or socializing hashtags, computed at different values of log followers (averaged over x). Uncertainty bands are 95% confidence intervals for the mean, computed by bootstrapping.

Besides engagement tactics, firms also design campaigns featuring different types of content, typically based on the specific product or promotion they are advertising. To understand how the FEI curve varies by the actual topic of the video, we first classify videos using a k-means clustering of the associated hashtags (see the Web Appendix for more details). We find five broad content topics: life (e.g., #WeekendVibes, #GoodMorning), holidays (e.g., #Christmas2020, #HappyHolidays), skills (e.g., #Yoga101, #InkDrawing), food (e.g., #HomeCooked, #HealthyCooking), and

gaming (e.g., #GamerGoals, #GamingTikTok). Again, we find that the average FEI curve varies substantially based on topic. We plot the average FEI curve by topic in Figure 8.

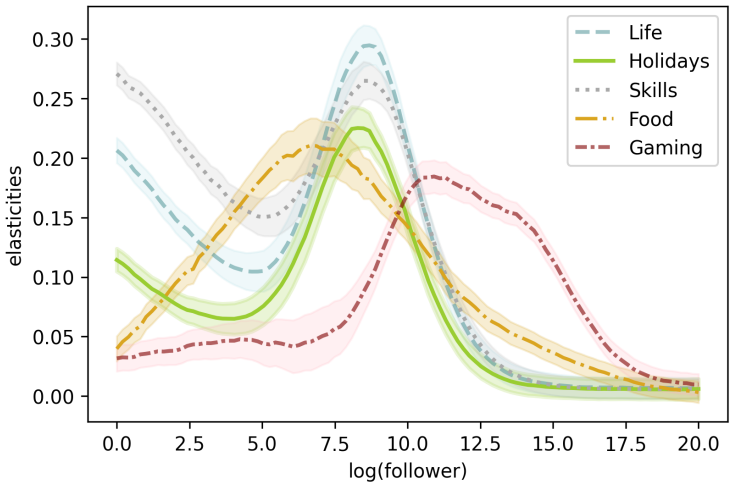


Figure 8: Average FEI by Content Topics
The average FEI for videos that were categorized by their hashtag as belonging to one of five content topics, reflecting what the videos are about at a high level. The FEI is computed at different values of log followers, averaged over x within the topic. Uncertainty bands are 95% confidence intervals for the mean, computed by bootstrapping.

We find that, while FEI always peaks somewhere in the mid-range of followers, the timing of this peak, and the overall magnitude of the elasticity, varies significantly by content topic. Holiday posts, for instance, have a generally lower elasticity than all other video types, suggesting that, overall, the returns on holiday-themed influencer campaigns may be lower than other types. In terms of the shapes of the curves, we see that, while food and gaming topics are inverted U-shaped, their peaks are substantially different: gaming videos peak much later than food videos, suggesting there may be higher returns to sponsoring more popular influencers in the context of gaming. Moreover, we see that for the other three topics – life, holidays, and skills – the curves exhibit what we call an inverted N-shape: they have an early elasticity peak, followed by a second peak in middle levels of log followers. This shape suggests that, for these video types, investing in small tier influencers may be especially effective. One plausible explanation, especially in the life and skills categories, is that smaller influencers feel more authentic, leading to higher elasticities for low follower counts.

4.2 Patterns Across Sponsored Campaigns

Until now, we have focused on establishing general patterns of FEI across video types, regardless of whether those videos were part of a sponsored challenge or not. Now, we zero in on just the sponsored challenges in our data, to understand how FEI varied across them. Recall that our data contained 30 sponsored challenges. Because the content and structure of each of these campaigns was different, we may expect to see different FEI curves. Thus, for each campaign, we estimate an FEI curve, based on the x learned from the official videos posted under that challenge. To simplify our analysis, after learning the FEI curve for each video, we compute the average FEI over six ranges, corresponding to six tiers of influencers: sub-nano (fewer than 1,000 followers), nano (1-10K followers), micro (10-50K), mid-tier (50-500K), macro (0.5-1M), or mega (more than 1 million followers). We report these average FEIs for each campaign in Table 4.

As seen in Table 4, we observe three recurring patterns in how the FEI, and, by implication, the counterfactual predictions, vary across campaigns: the FEI curves are either first increasing and then decreasing, which implies an S-shaped counterfactual prediction curve (like we observed on average in Figure 6); relatively constant, corresponding to a linear counterfactual prediction curve; or monotonically decreasing, corresponding to a concave counterfactual prediction curve. For example, in Figure 9, we plot the FEI and counterfactual prediction curves for Walmart's #UnwrapTheDeals campaign, which we previously discussed in the introduction (see Figure 1), and whose curves are similar to the average shapes shown previously in Figure 6. On the plot, for reference, we also label regions of the x-axis corresponding to the six tiers of influencers. In contrast, in Figures 10 and 11, we plot two examples from Apple and Dettol (the latter of which was the other video in Figure 1). Relative to the Walmart video, the Apple video (which was a workout tutorial called #CloseYourRings) has a counterfactual prediction curve that is very concave, suggesting that the highest marginal effectiveness comes from smaller influencers. Said differently, the concave shape suggests that additional followers become less and less effective in reaching consumers. In contrast, the Dettol video's counterfactual prediction curve is more linear, suggesting the marginal effectiveness is more consistent across influencer tiers. Even at high levels of popularity, this shape suggests there are still sizable reach gains possible with increasing follower sizes. These differences will have major implications for optimal influencer partnerships, as we

	<Nano	Nano	Micro	Mid-tier	Macro	Mega
S-shaped						
#UnwrapTheDeals	0.21	0.42	0.38	0.19	0.11	0.05
#WhatsYourPower	0.22	0.45	0.42	0.27	0.10	0.06
#VideoSnapChallenge	0.16	0.54	0.62	0.29	0.13	0.05
#GetCrocd	0.14	0.48	0.66	0.30	0.14	0.05
Linear						
#CaliStarChallenge	0.21	0.24	0.20	0.28	0.26	0.23
#HandWashChallenge	0.25	0.36	0.32	0.27	0.30	0.20
#MoreHappyDenimDance	0.18	0.24	0.20	0.21	0.25	0.22
#ScoobDance	0.15	0.17	0.14	0.22	0.20	0.16
#ASOSFashunWeek	0.32	0.39	0.33	0.31	0.34	0.28
#TheSplashDance	0.25	0.27	0.31	0.30	0.32	0.34
#ReadySETgo	0.29	0.34	0.29	0.37	0.30	0.26
#LetsFaceIt	0.27	0.32	0.30	0.37	0.33	0.28
#DoPacSun	0.35	0.29	0.27	0.28	0.22	0.20
#MerryBOSSmas	0.16	0.20	0.14	0.18	0.12	0.10
#MoodFlip	0.21	0.27	0.22	0.30	0.28	0.25
#ThisIsBliss	0.26	0.34	0.32	0.36	0.34	0.32
#UptheBeat	0.19	0.26	0.28	0.25	0.25	0.22
#katespadenyhappydance	0.12	0.18	0.22	0.22	0.20	0.14
#PerfectAsIAm	0.22	0.34	0.28	0.28	0.32	0.30
#exprESSIEyourself	0.14	0.20	0.19	0.22	0.23	0.22
Concave						
#CloseYourRings	0.43	0.24	0.24	0.12	0.11	0.05
#ItWasntMe	0.38	0.22	0.20	0.17	0.10	0.04
#HeinzHalloween	0.44	0.28	0.23	0.15	0.09	0.05
#StrictlyCurl	0.62	0.27	0.18	0.13	0.11	0.06
#MONCLERBUBBLEUP	0.35	0.18	0.18	0.13	0.09	0.03
#ShowUpShowOff	0.68	0.33	0.24	0.16	0.12	0.05
#GotMilkChallenge	0.53	0.42	0.22	0.18	0.09	0.04
#GoForTheHandful	0.66	0.27	0.19	0.17	0.11	0.05
#MicellarRewind	0.39	0.22	0.20	0.14	0.10	0.03
#CancelTheNoise	0.45	0.28	0.21	0.15	0.12	0.05

Table 4: Campaign FEI by Influencer Tiers
For each sponsored challenge, we compute the average of the FEI curve of that challenge over six intervals, corresponding to different influencer tiers. We then grouped these curves into three buckets, based on the implied shape of the corresponding counterfactual prediction curve: S-shaped, linear, and concave.

discuss in a later subsection.

Before we discuss the implications for optimal influencer selection, we first want to understand if there are any factors that seem predictive of which types of campaigns yield each of the three curve types. To that end, we examine the actual content of each campaign, including its name, description, and the types of videos posted under that challenge. We give examples of campaigns in each curve type category in Table 5. We find that the S-shaped, linear, and concave growth tightly correspond to three salient campaign attributes: campaigns based on adding *special effects* to videos tend to exhibit S-shaped curves; challenges that encourage *self-expression* via either dancing or outfit change tend to exhibit linear curves; and *product demonstration* challenges tend to exhibit concave curves.

There are several plausible explanations for these patterns. The first relates to the audience size, or demand, for particular types of campaigns: product demonstration videos, which exhibit a concave impression growth curve, can only attract an audience that is interested in their products, which is a smaller niche than, say, the community of TikTok users interested in dancing and outfit change videos. As a result, smaller influencers yield a higher marginal effectiveness in reaching that niche audience than larger influencers. For special effects videos, the stronger S-shaped pattern suggests a more prominent thresholding effect: a special effect will only catch on if a sufficiently popular influencer posts it.

Another possible explanation is follower quality. While our model implicitly assumes all followers are equally interested in the content posted by an influencer, that is not necessarily true: an influencer's early followers may be more interested than those who "hop on the bandwagon" once the influencer is already well established. The concave shape, then, may be a result of this higher level of interest for early followers in the case of product videos, while more linear shapes suggest that all followers are equally interested.

4.3 Connections to the Advertising Literature

Although we have described our findings so far in the context of managers of influencer marketing campaigns, influencer marketing can be viewed as a type of advertising on social media. Thus, our findings can also speak to the extant literature on advertising response, in the context of influencer

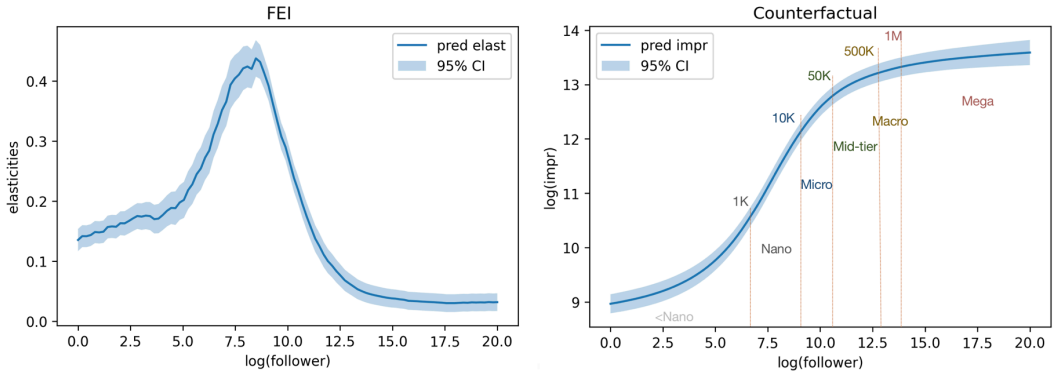


Figure 9: Walmart's #UnboxTheDeals FEI and Counterfactual Prediction Curves
At left, the FEI curve, and at right, the counterfactual prediction curve for impressions of Walmart's #UnwrapTheDeals video, given the (log) number of followers. The vertical lines in the right panel delineate the six influencer tiers on the x-axis.

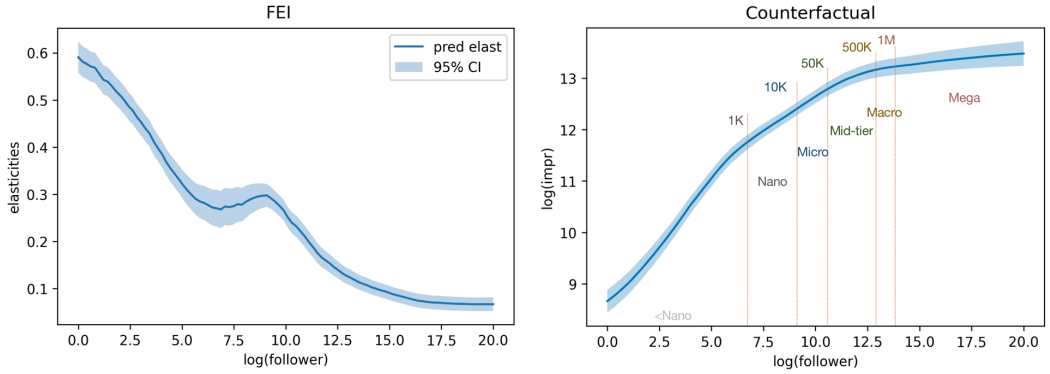


Figure 10: Apple's #CloseYourRings FEI and Counterfactual Prediction Curves
At left, the FEI curve, and at right, the counterfactual prediction curve for impressions of Apple's #CloseYourRings video, given the (log) number of followers. The vertical lines in the right panel delineate the six influencer tiers on the x-axis.

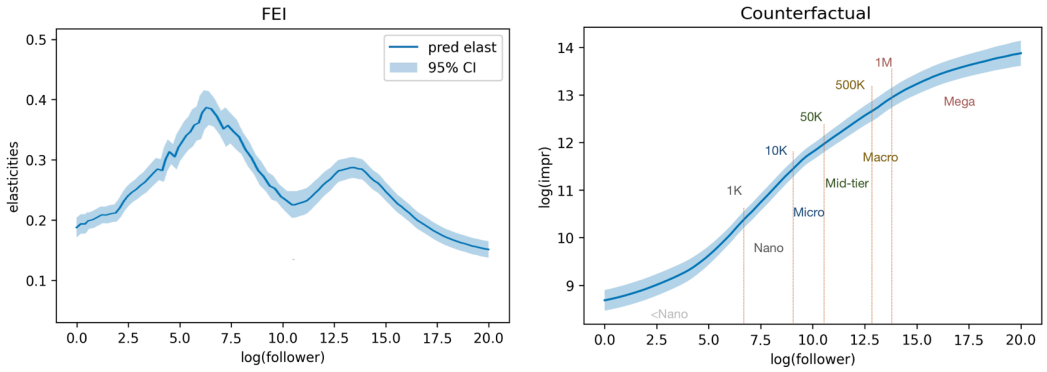


Figure 11: Dettol's #HandWashChallenge FEI and Counterfactual Prediction Curves
At left, the FEI curve, and at right, the counterfactual prediction curve for impressions of Dettol's #HandWashChallenge video, given the (log) number of followers. The vertical lines in the right panel delineate the six influencer tiers on the x-axis.

	Description	Frames
S-shaped		
#UnwrapTheDeals	Ready to #UnwrapTheDeals? Use our **#UnwrapTheDeals effect**, reveal your deal, then post with the hashtags #UnwrapTheDeals and #contest for a chance to win.	
#WhatsYourPower	If you had the opportunity to have **superpowers** for five minutes, but you couldn't choose what they were…would you take it? Watch Project Power, only on Netflix.	
#GetCrocd	It's time to #GetCrocd. Use our **Crocs effect** to try on three different pairs (with jibbitz!) and strut your stuff.	
Linear		
#UptheBeat	Let's get down and #UpTheBeat. Challenge: throw on your favorite FILA look, learn this **dance** and put your own spin on it.	
#ThisIsBliss	Push away any self-doubt and celebrate with your blissful **dance**!	
#DoPacSun	You know how to dress, and you know how to **do transitions**. You did that. Now #DoPacSun. Show us your favorite PacSun looks, and let's see who owns this sound.	
Concave		
#StrictlyCurls	Curly, Wavy and Straight hair crews, we want to see you **take your texture to the next level** with some TikTok flair.	
#MicellarRewind	**Rewind your routine** with Garnier Micellar Water! Show off your Micellar transformation in 3,2,1….	
#CancelTheNoise	This holiday season, it's time to tap into the joy of Bose QuietComfort Earbuds. Show us how you **double tap to #CancelTheNoise** and feel it all.	

Table 5: Campaign Descriptions by Curve Type
For each of the curve types, we show four examples of campaigns with that curve type, including the hashtag name, a description of the campaign, and an example video.

marketing. We now make this connection more explicit.

The literature on advertising has extensively explored question of marginal effectiveness of advertising with respect to exposure. In that literature, there has been some debate as to whether advertising response is S-shaped or concave, and whether a threshold effect exists, wherein an ad is only effective if its exposure exceeds a certain level (Rao and Miller, 1975; Johansson, 1979; Cannon et al., 2002; Vakratsas et al., 2004). Distinct from this literature, our response curves use impressions, rather than sales, as the dependent variable, and follower size, rather than advertising expenditure, as the explanatory variable. Nonetheless, there are parallels: since influencers are typically paid based on their follower sizes, and since a given follower size cannot *guarantee* a given number of impressions, follower size plays a similar role to advertising expenditure. Likewise, while impressions are not the same as sales, they are often a metric of primary concern in influencer marketing, and can be assumed to be proportional to sales.

Building on these equivalencies, our results suggest that, in the context of influencer advertising, an S-shaped response curve indeed can exist, and is quite prevalent, but only for certain types of content. In addition, if there is a threshold effect, it disappears at an extremely early point (i.e., at less than 1,000 followers). Generalizing beyond our context, these findings suggest that there may be two reasons for the limited empirical evidence of S-shaped responses to advertising. First, threshold effects may not exist in the limited product classes for which data are available in past studies. Second, while the range of followers of social media users can range from zero to the millions, more traditional advertising is typically not done at extremely low levels. Given we find a threshold exists at only very small numbers of followers, its likely that such an effect would be impossible to observe with traditional advertising data. We also identify concave and linear shaped response curves for different content, the latter of which has rarely been documented in the advertising literature. It indicates that for some content, saturation of the potential audience is yet to come, making influencers (individuals with large followings) more important in expanding the potential of some campaigns to go viral. That we observe these response curves on TikTok, but not historically in advertising may be attributable to the less mature content market on social media that still experiences explosive growth in attention.

4.4 Optimal Influencer Selection

Finally, now that we have established some patterns in when different patterns in FEI emerge, we turn to the most important implication of FEI: determining which influencers a firm should sponsor. Just as price elasticity of demand is crucial for determining optimal pricing strategies, the follower elasticity of impressions is crucial for determining optimal influencer partnerships. Our video-level FEI curves can guide firms in choosing an optimal popularity level for an influencer to collaborate with for a specific campaign. Doing so requires three ingredients: first, the features of the desired video, x, from which we can compute the predicted FEI curve for that video; second, an assumption about how impressions translate into revenue; and third, the cost structure of the sponsorship. With these three ingredients, we can use the FEI, together with how impressions translate to revenue, to compute an estimate of the marginal revenue associated with an additional follower. Then, from the cost structure, we can estimate a marginal cost, again in terms of an additional follower. Setting the two equal yields the first order conditions for optimal influencer selection.

To illustrate, we focus on the case study of Walmart's #UnwrapTheDeals campaign that was shown in the right panel of Figure 1, and discussed again Figure 9. The focal question is, what level of popularity should Walmart target for an influencer to promote this campaign? To operationalize "this campaign," we use the features from the actual video depicted in Figure 1. With this x, we then compute the FEI curve, and the counterfactual prediction curve, which we previously displayed in Figure 9. As shown in the plot, the impression of the same Walmart video is predicted to range from several thousands to more than 1 million when posted by different influencers, indicating that follower size is indeed crucial in driving video impressions. Generally speaking, more followers lead to more video impressions. However, the speed at which the number of impressions grow differs across the influencer tiers, again following a similar effect as before: the highest marginal return comes from the middle. Depending, then, on the marginal revenue with respect to followers, partnering with the most popular influencer may be suboptimal.

To actually determine an optimal partnership based on this curve, we need to make some assumptions about the way impressions generate value for Walmart, as well as the costs of the partnership. For this study, we assume a \$value/impression rate of \$0.02, which was set to match

the average cost-per-view of a video ad on Youtube.[13] On the cost side, firms typically pay influencers based on their follower size.[14] Assuming that the payment is linear in the number of followers, the marginal cost of sponsoring an influencer for an ad post is constant. Following industry norms, we assume a pay rate at \$10/1K followers.[15]. Based on this, we find that the optimal influencer for our Walmart example is a mid-tier influencer with around 150K followers. Our analysis suggests that Walmart's maximum profit (\$8,225) from paying the suggested influencer is 37% more than paying the best micro influencer with 50K followers and is 700% more than paying a mega influencer with 1.5M followers, which was Walmart's choice in reality.

While the previous analysis was based on a single set of assumptions, the benefit of thinking about optimal selection through FEI is that it is generalizable to any assumed cost and revenue structures. Thus, in Figure 12, we analyze under what conditions a given follower size would make sense, in terms costs per thousand followers, and revenue per impression. In each cell, we label the follower size (in thousands of followers) for the optimal influencer. Note that in the context of the Walmart campaign, influencers with millions of followers (suggested at the left bottom corner of the heatmap) will be justified only if either their payment rate is low, and (or) the firm places a high value on an impression on the TikTok platform.

Finally, we consider how the different curve shapes described in the previous subsection suggest different optimal follower sizes. Following the logic described previously, Walmart's impression curve is S-shaped, suggesting increasing followers in the mid-range yields the most "bang for the buck." For concave curves, on the other hand, like Apple's #CloseYourRings campaign, increasing followers on the lower end of the distribution is more effective. This suggests that, all else equal, a smaller number of followers may be more optimal. Dettol's #HandWashChallenge, on the other hand, had an almost linear impression growth curve: linear curves suggest that the marginal returns by increasing followers remain roughly constant, suggesting a higher number of followers may be more optimal. Indeed, using our FEI-based logic, with the same assumptions as above, the optimal number of followers for Apple and Dettol follow this logic: we compute the optimal follower size for the Apple campaign is around 100K, substantially less than Walmart, while the optimal size for Dettol is over twice that, 225K, substantially more than Walmart.

[13] https://influencermarketinghub.com/how-much-do-youtube-ads-cost/

[14] https://influencermarketinghub.com/influencer-rates/

[15] https://nealschaffer.com/how-much-to-pay-influencer/

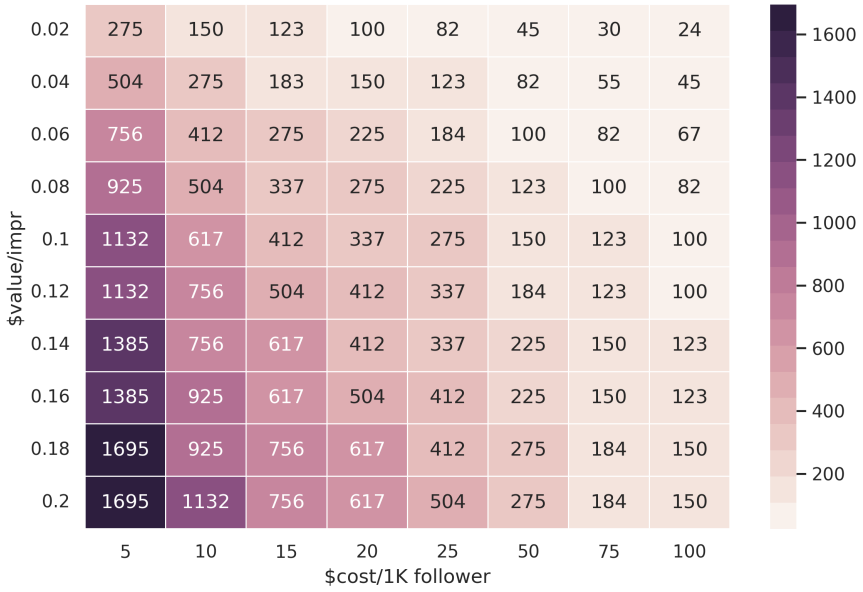


Figure 12: Walmart's Optimal Follower Size By Condition
The heatmap shows the optimal number of followers (in thousands) Walmart should target based on the cost per thousand followers on the x-axis, and revenue per impression on the y-axis.

5 Conclusion

In this work, we have examined the relationship between an influencer's follower size and the reach of their videos on the short video platform TikTok. Given the explosive popularity of TikTok for influencer marketing, understanding this relationship is an essential ingredient in firms' decisions about influencer marketing. Since cost structures in influencer marketing typically depend on the popularity of the influencer, firms must understand how this popularity actually translates into reach, in terms of the number of video impressions, to optimize their TikTok influencer campaigns. In this vein, our work makes several contributions: first, we propose a combination of representation learning and causal machine learning methods, which can be used to carefully quantify the causal effect of popularity on reach (or, in our case, number of followers on impressions). From this analysis, we derive the central quantity of follower elasticity of impressions, FEI, which can be used to decide the optimal popularity level of their influencer partners. Through our application of the framework to TikTok, we have demonstrated the rich heterogeneity in FEI across many variables of interest to firms, and illustrated how differences in FEI affect optimal decision making. Finally, our findings speak to several on-going academic debates about response to advertising and optimal advertising strategies. While our empirical setting and specific find-

ings are restricted to TikTok, our methodological framework is generalizable to other social media platforms.

In short, without careful causal analysis and consideration of the FEI, our results suggest that firms may erroneously err toward thinking that more popular TikTok influencers are better. Instead, we show that, counter to what might be concluded from simple correlation analyses, on average, the highest returns come from considering influencers of middling levels of popularity. That is, our follower elasticity metric often exhibits an inverted U-shape with respect to influencer popularity. Moreover, our findings suggest the content of a campaign plays a huge role in how popular of an influencer should be optimally sponsored, and more generally in how a video's eventual impressions respond to its creator's popularity. We show that optimal influencer strategies vary in predictable ways, based on how the firm is trying to engage with customers, what their campaigns are about in terms of content, and how exactly their campaigns are structured, in terms of the kinds of videos they encourage customers to construct. Based on how campaigns vary along these dimensions, their follower elasticity curves may exhibit a variety of shapes, suggesting different optimal partnerships. By characterizing these general patterns along interpretable, actionable factors, we give managers a framework to think about what makes influencer campaigns successful, beyond the simple case study analyses that are common pointed to in practice.[16]

While our work offers one way to systematically evaluate influencer partnerships, it is not totally comprehensive. Here, we only considered how influencers vary on a single dimension: popularity. In reality, influencers may also vary in other ways, including on what type of content they typically post, and how well that content matches the brand. We have also only analyzed the case where a firm is collaborating with a single influencer. In the case they wish to sponsor multiple influencers, there may be more complexities to consider, although the concept of FEI will still be important. In general, we hope future work will build on the methods and metrics developed in our work to explore these questions in more depth. Influencer marketing is a complex topic that is of central interest to modern marketers; we hope our work makes an initial contribution to the study of the topic that can be built on by future research.

[16]For example: https://www.tiktok.com/business/en-US/blog/branded-hashtag-challenge-harness-the-power-of-participation

References

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455.

Athey, S., Tibshirani, J., and Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178.

Berger, J. and Milkman, K. L. (2012). What makes online content viral? *Journal of marketing research*, 49(2):192–205.

Bhattacharya, S., Nojavanasghari, B., Chen, T., Liu, D., Chang, S.-F., and Shah, M. (2013). Towards a comprehensive computational model foraesthetic assessment of videos. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 361–364.

Borth, D., Ji, R., Chen, T., Breuel, T., and Chang, S.-F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232.

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

Cannon, H. M., Leckenby, J. D., and Abernethy, A. (2002). Beyond effective frequency: Evaluating media schedules using frequency value planning. *Journal of Advertising Research*, 42(6):33–46.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., and Newey, W. (2017). Double/debiased/neyman machine learning of treatment effects. *American Economic Review*, 107(5):261–65.

Dekimpe, M. G. and Hanssens, D. M. (2007). Advertising response models. *Handbook of advertising*, pages 247–263.

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

Dolan, R., Conduit, J., Frethey-Bentham, C., Fahy, J., and Goodman, S. (2019). Social media engagement behavior: A framework for engaging customers through social media content. *European Journal of Marketing*.

Erhan, D., Courville, A., Bengio, Y., and Vincent, P. (2010). Why does unsupervised pre-training help deep learning? In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 201–208. JMLR Workshop and Conference Proceedings.

Farrell, M. H., Liang, T., and Misra, S. (2021). Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213.

Gelman, A., Goodrich, B., Gabry, J., and Vehtari, A. (2019). R-squared for bayesian regression models. *The American Statistician*.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. (2017). Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE.

Goldenberg, J., Han, S., Lehmann, D. R., and Hong, J. W. (2009). The role of hubs in the adoption process. *Journal of marketing*, 73(2):1–13.

Guo, T., Sriram, S., and Manchanda, P. (2021). The effect of information disclosure on industry payments to physicians. *Journal of Marketing Research*, 58(1):115–140.

Hartford, J., Lewis, G., Leyton-Brown, K., and Taddy, M. (2017). Deep iv: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Hong, L., Dan, O., and Davison, B. D. (2011). Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World wide web*, pages 57–58.

Hsu, W.-N., Zhang, Y., and Glass, J. (2017). Learning latent representations for speech generation and transformation. *arXiv preprint arXiv:1704.04222*.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7(1):443–470.

Johansson, J. K. (1979). Advertising and the s-curve: A new approach. *Journal of Marketing Research*, 16(3):346–354.

Kannan, P. et al. (2017). Digital marketing: A framework, review and research agenda. *International journal of research in marketing*, 34(1):22–45.

Katona, Z., Zubcsek, P. P., and Sarvary, M. (2011). Network effects and personal influences: The diffusion of an online social network. *Journal of marketing research*, 48(3):425–443.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Li, C., Rudin, C., and McCormick, T. H. (2022). Rethinking nonlinear instrumental variable models through prediction validity. *Journal of Machine Learning Research*, 23(96):1–55.

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

Liu-Thompkins, Y. (2012). Seeding viral content: The role of message and network factors. *Journal of advertising research*, 52(4):465–478.

Liu-Thompkins, Y. and Rogerson, M. (2012). Rising to stardom: An empirical investigation of the diffusion of user-generated content. *Journal of Interactive Marketing*, 26(2):71–82.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578.

Rajaram, P. and Manchanda, P. (2020). Video influencers: Unboxing the mystique. *arXiv preprint arXiv:2012.12311*.

Rao, A. G. and Miller, P. B. (1975). Advertising-sales response functions. *Journal of Advertising Research*, 15(2):7–15.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Susarla, A., Oh, J.-H., and Tan, Y. (2012). Social networks and the diffusion of user-generated content: Evidence from youtube. *Information systems research*, 23(1):23–41.

Vakratsas, D., Feinberg, F. M., Bass, F. M., and Kalyanaram, G. (2004). The shape of advertising response functions revisited: A model of dynamic probabilistic thresholds. *Marketing Science*, 23(1):109–119.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wei, Y., Wang, X., Guan, W., Nie, L., Lin, Z., and Chen, B. (2019). Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing*, 29:1–14.

Yang, J., Zhang, J., and Zhang, Y. (2021). First law of motion: Influencer video advertising on tiktok. *Available at SSRN 3815124*.

Yao, R., Liu, C., Zhang, L., and Peng, P. (2019). Unsupervised anomaly detection using variational auto-encoder based feature extraction. In *2019 IEEE International Conference on Prognostics and Health Management (ICPHM)*, pages 1–7. IEEE.

Zantedeschi, D., Feit, E. M., and Bradlow, E. T. (2017). Measuring multichannel advertising response. *Management Science*, 63(8):2706–2728.

Zha, S., Luisier, F., Andrews, W., Srivastava, N., and Salakhutdinov, R. (2015). Exploiting image-trained cnn architectures for unconstrained video classification. *arXiv preprint arXiv:1503.04144*.

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

Zhang, Y., Moe, W. W., and Schweidel, D. A. (2017). Modeling the role of message content and influencers in social media rebroadcasting. *International Journal of Research in Marketing*, 34(1):100–119.

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., and He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76.

Web Appendix

Web Appendix A: Post Statistics

For each post, our post statistics include metadata information about 1) the hashtags used by the post, 2) the ranking and age of the post within the hashtag, 3) the sponsorship status of the post, and 4) the post's creator. We now describe each of these in more detail and present their summary statistics below.

Hashtags Each post may contain multiple hashtags, which influence how widely a video is seen. One such hashtag is #fyp, which stands for "For You Page." The For You page (FYP) is TikTok's personalized landing page, where the platform recommends videos to their users based on its internal algorithm. It is widely believed that including #fyp on a post increases the likelihood of a post being featured on the FYP. Thus, we include an indicator variable to capture whether a post contains the #fyp hashtag ("Has FYP"). We also track the total number of hashtags used in a post ("Num hashtags") and among them how many are *trending* ("Num trending").

Rank and Age On the Discover page, under each hashtag, there is a list of videos. The videos that are listed towards the top of the page get more exposure than those lower down. We track a video's rank under each hashtag on the Discover page ("Ranking"), and whether the hashtag was trending ("Is trending"). Finally, for each day a post is tracked, we record how old the post is ("Video age") and how old the hashtag is ("Hashtag age") to represent their relative trendiness.

Sponsorship For each post, we check if it is sponsored by a firm ("Is sponsored") or organic.

Creator For each creator, we check if he/she is a verified user ("Is verified") as a general measure of their speciality from various aspects.

Variable	N	Mean	Std. Dev.	Min	Median	Max
Num Hashtags	518,303	6.662	3.275	0	6	47
Num Trending	518,303	1.078	0.853	0	1	10
Has FYP	518,303	0.604	0.489	0	1	1
Hashtag Age	518,303	6.003	13.316	0	1	148
Is Trending	518,303	0.767	0.423	0	1	1
Ranking	518,303	1029	586.495	1	1,037	2,000
Video Age	518,303	35.212	102.947	0	1.327	2,015
Is sponsored	518,303	0.001	0.031	0	0	1
Is verified	518,303	0.003	0.171	0	0	1

Table 6: Summary Statistics of the Post Statistics

Web Appendix B: VS-VAE Architecture Details

The text encoder consists of a Bi-LSTM layer with dimension size 64 for the three text sources and a fully connected layer of the same size 64 for the concatenated output from the three Bi-LSTM layers. The image encoder consists of one LSTM layer with dimension size 4096 for the sequential VGG-19 features and three fully connected layers of sizes 4096, 1024 and 512 for the final image vector including the LSTM output of the VGG-19 features, the object vector and the visual sentiment vector. The audio and editing encoders consist of one fully connected layer of size 64 and 16, respectively. The joint encoder takes the concatenation of the encoded features of size 656 from the four separate encoders as input and further compresses it into size 256 through one fully connected layer. The decoders consists of layers with the same dimensions yet inverted as those of the encoders. For all fully connected layers in both encoders and decoders, we use the Rectified Linear Unit (ReLU) as the nonlinear activation function.

We experimented with batch sizes 16, 32, 64, 128, 256, and found 32 to be optimal for the training of our VS-VAE. The model was trained for 300 epochs with early stopping and a learning rate of 1e-5. To prevent overfitting, we use L2-regularization on the weights of our model. We experimented with weight penalties of 0, 0.05, 0.1, 0.3, and 0.6 and found 0.05 to be optimal for both the encoder and decoder. To minimize our VS-VAE loss, we use Adam as the optimizer. As mentioned in the data section, our video data have around 1500-2000 videos under each of the hashtags. In total, we have more than 0.5M unique videos spanning over all hashtags. They are randomly split by hashtag into three parts: the training set, the validation set and the test set, at the ratio of 8:1:1. In other words, they are split randomly but the ratio is preserved within each hashtag. This strategy ensures, at least from the data perspective, that all impactful, recurring content attributes resulting from belonging to a specific hashtag on video impression growth can be well-captured, even under some imbalance in the video volume across hashtags.

Web Appendix C: Deep IV Implementation Details

As in Hartford et al. (2017), we train the networks in both stages of Deep IV via stochastic gradient descent. To select the best hyperparameters, we perform out-of-sample validation for both stages by simply evaluating their respective losses on held out data. Each stage is evaluated sequentially, with the second stage validation based on the best possible network from the first stage. In the end, for the model for F we find the optimal architecture is a mixture of 10 Gaussian distributions, with 3 fully connected layers with sizes 128, 64 and 32 for the network modeling their parameters. For the model of h, we find an optimal structure of 4 fully connected layers, with size 128, 64, 32 and 1, respectively. For all layers, we use the Rectified Linear Unit (ReLU) as the activation function. The model is trained on batches of size 256 for 100 epochs with early stopping and a learning rate of 1e-5. To prevent overfitting, we set the dropout rate to be 0.2.

Web Appendix D: Deriving the Hashtag Topics

To classify each of our hashtags into a content topic, we learn embeddings of the words within that hashtag, then use those embeddings to derive a representation of the overall meaning of the hashtag, which we then cluster across hashtags. Hashtags are typically concatenations of common words. Hence, we first develop an algorithm that automatically segments each hashtag into separate words. The algorithm works by subsetting the overall hashtag name into strings which can be matched to a dictionary. Then, the words that are found are matched to pre-trained GloVe embeddings. This yields 50-dimensional word embeddings for each segmented word. To combine them into a single embedding, we use average pooling over all embeddings in the hashtag name. As an example, consider the hashtag "#HomeOffice": we first segment it into "Home" and "Office" and take the average of the two word embeddings as the semantic representation for "#HomeOffice". Consequently, each hashtag is represented by a single 50-dimensional vector. We then identify five topics among all hashtags that each has at least 15 exemplars in our dataset based on the k-means clustering of their semantic representations. Matching the hashtag closest to each of the five cluster centers to the most popular content categories suggested on TikTok,[17] we summarize the five topics as 1) life, 2) holidays, 3) skills, 4) food, and 5) gaming.

[17] https://www.statista.com/statistics/1130988/most-popular-categories-tiktok-worldwide-hashtag-views/

Web Appendix E: Extracting Content Features from TikTok Posts

From each post on TikTok, we extract features from four modalities: (1) textual features, from both the video and the caption attached below; (2) image features, which are learned from the image frames of the video; (3) audio features, which capture features of the video's sound; and (4) editing features, which capture how creators edit their videos on the platform. While the first three are standard in video analysis, the fourth is a modality unique to TikTok. TikTok features numerous built-in editing tools for videos, which allow creators to add different effects to their videos (e.g. visual filters, musical and animal voice effects). Given the centrality of these features to TikTok posts, we create an innovative set of features to measure their presence in a post. We now describe each of these feature sets in more detail.

Text Our textual features are based on three sources: the video's caption, any text overlaid on the videos themselves (referred to as *stickers*), and text capturing the words spoken in any video voice-overs. Both the video description and sticker text are readily extracted from the TikTok post. For the voice-over, we first identify if the background music of a video belongs to any sound types related to speaking (e.g. speech, conversation or monologue) using YAMNet, a pre-trained audio classification model (Gemmeke et al., 2017; Hershey et al., 2017). If so, we then employ Google's speech-to-text API to convert the audio to words.[18] Across all three text modalities, we further process the raw text into word embeddings, leveraging the pre-trained Word2Vec model (Mikolov et al., 2013). Hence, our final set of textual features is a collection of word embeddings, across these three sources.

Image The visual component of video data consists of sequences of image frames. In our analysis, rather than extracting and analyzing each individual image frame, of which there may be thousands, we extract image frames in five second intervals. By doing so, we greatly reduce the volume of data, while still maintaining a relatively rich set of image frames.[19] We apply two pre-trained architectures to process these video frames into more meaningful features.

The first architecture is the popular VGG-19 model, a 19-layered deep neural network origi-

[18] https://cloud.google.com/speech-to-text

[19] A TikTok video is normally 3-60 seconds long with 30-60 frames per second.

nally trained to classify 1,000 unique classes of objects in images (Simonyan and Zisserman, 2014). For each of our image frames, we feed the frame through VGG-19, and extract the second-to-last (4,096-dimensional) fully connected layer. Intuitively, the layers of VGG-19 learn representations of image content at increasing degrees of abstraction. Thus, the later layers of the model yield a high-level representation of the image content. A robust set of prior studies has shown that the final layers of the VGG-19 architecture can serve as the basis for a variety of other prediction tasks (e.g. Wei et al., 2019; Zha et al., 2015), exactly because they form an abstract representation of the image content, which is what we take advantage of here. In addition to the 4,096-dimensional representation, we also include as visual features the actual 1000-dimensional object probabilities returned by VGG-19. These probabilities capture whether one of the 1,000 classes of objects VGG-19 was trained on are present in a shot, giving us a more concrete understanding of what, exactly, was in each frame of the video.

The second architecture is a pre-trained convolutional neural network model trained on the SentiBank data (Borth et al., 2013). The SentiBank dataset is a set of tagged images used for "visual sentiment" analysis. The model provides probabilities over two-word (adjective-noun) image concepts, like "creepy house" or "happy dog." In contrast to simple sentiment schemes like positive-negative, the SentiBank ontology is richer and captures many of the concepts actually used in social media posts. The number of visual concepts is quite large (2,089), and the output of the SentiBank is a 2,089-length vector of probabilities.

In sum, for each image frame, our set of features comprise of three vectors: the 4,096-dimensional VGG-19 representation, which captures a holistic representation of the image; (2) the 1,000-dimensional VGG-19 object class probabilities, which captures a wide range of potential objects that might be present in the image; and (3) the 2,089-dimensional SentiBank visual sentiment probability vector, which captures emotional features of the image. In the end, we performed a max pooling over the image frames to generate a single object vector and visual sentiment vector for maximal inclusiveness. Together with the matrix of VGG-19 features with dimension #frames$\times$4096, they summarize all the extracted visual features for each video.

Audio The other key aspect of video data is the audio track. In fact, audio is a central component of TikTok as it often provides a unifying link between videos within a hashtag. Background audio

on TikTok can be selected from TikTok's music library, from music clips used in other videos, or uploaded as "original sound" by the creator. Our audio features try to capture both the features of the sound itself, as well as, how the sound connects to other videos on the platform. Specifically, we collect three types of audio features:

- **Universality:** A single number from 0 to 1 that captures how many other videos within a hashtag use the same audio.
- **Sound classes:** We feed the audio through YAMNet (Gemmeke et al., 2017; Hershey et al., 2017), a deep neural net that predicts 521 audio event classes, based on the AudioSet-YouTube data. These features include various sound bytes that characterize laughter, conversation, and different musical instruments. The specific feature we use is the 521-dimensional output probabilities, averaged over the entire sound track.
- **Acoustic features:** Finally, we extract standard acoustic features, capturing the raw traits of the sound itself (e.g., intensity and pitch).

In sum, the audio for video is summarized by a 543-dimensional representation.

Editing TikTok offers various tools that allow users to easily add special effects to their videos, including simple edits such as video length and speed, and more complex edits such as image filters. These tools are frequently used, and are often highly salient in the final videos posted on the platform. We summarize these features below:

- **Video length:** The actual length of the video, in seconds, which tends to correlate with video content. For instance, cooking-related videos can be either short (e.g., just show the finished product), or long (e.g., tutorial-style with all the steps).
- **Stickers:** Whether the videos have textual stickers added to them, and if so, how many, and how many words are in each.
- **Scene Length:** To quantity the overall speed of visual transitions of a video, we use the average scene length which might vary remarkably from video to video. Specifically, we identify the scene boundaries or transitions that occurs in a video by identifying the discontinuity between two consecutive image frames. We adopt the traditional intensity-based

scene detection algorithms that compare the sharpness of the changes in color histograms across consecutive frames. With the scene transitions, a video scene is identified as the content between each two consecutive transitions, and the length of a scene is then the length of the time period between the two transitions. Finally, we take the average of the scene lengths among all detected transitions.

- **Feature Variance:** Through editing, a creator could artificially create different content complexity to fit in different video genres. For example, in case of videos of news or story sharing with negligible background fluctuations, there are very small changes in both spatial and temporal domains. On the other extreme, in case of high-action content such as DIY projects, they are normally accompanied with rapid visual changes throughout the video. The amount of information with in and differentiating each frame, and thereby, the content complexity is fairly high. In the paper, we define both the spatial and temporal content complexity of a video as the average variance of the object/sentiment probabilities across its key frames and across the object/sentiment classes. Mathematically, for any video i,

$$\text{Spatial Content Complexity}_{i,k} = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \left(\frac{1}{|\mathcal{S}_k|} \sum_{i \in \mathcal{S}_k} (p_{ijt} - \overline{p_{it}})^2 \right)$$

$$\text{Temporal Content Complexity}_{i,k} = \frac{1}{|\mathcal{S}_k|} \sum_{j \in \mathcal{S}_k} \left(\frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} (p_{ijt} - \overline{p_{ij}})^2 \right)$$

  Here, j represents the class, $k \in \{\text{object, sentiment}\}$, $\mathcal{S}_k$ denotes the set of k, and $\mathcal{T}_i$ denotes the set of key image frames for video i. Then, p_{ijt} represents the probability of class j in frame t for video i, $\overline{p_{it}}$ represents the mean probability across all classes in frame t for video i, and $\overline{p_{ij}}$ represents the mean probability for class j across all image frames for video i.

- **Filters:** We cannot directly assess if a filter has been applied. However, we can measure the aesthetic quality of the image with metrics such as lightness and symmetry, which are affected by the use of image filters (Bhattacharya et al., 2013).

In total, we learn a 157-dimensional vector of editing features from the data.

Web Appendix F: Details of our Representation Learning Algorithm

While the feature extraction process helps structure the raw data, the dimensionality of the feature space it is still massive. To make these features more amenable for use within the deep IV framework, while also accounting for synergies and correlations across the different modalities, we build a representation learning model that embeds TikTok posts in a lower dimensional space. Specifically, our framework is based on the variational autoencoder (VAE), a deep generative model used previously for learning representations of multimodal data (Dew et al., 2022). It takes as input all of the content-related features, and returns a vector representation, $\boldsymbol{r}$ that captures the essence of the content, such that given $\boldsymbol{r}$, the original content could be reconstructed. Before describing our specific model, we briefly describe variational autoencoders.

F.1 Variational Autoencoders

The variational autoencoder is a type of autoencoder, which is a machine learning model that has two key components: an encoder that compresses the data to a dense vector representation and a decoder that reconstructs the original data from that representation. The variational autoencoder is a probabilistic variant of this framework, where the generative process of the observed data, $\mathbf{x}_i$ for observation i, is modeled as a function of latent lower-dimensional vector representations, $\mathbf{z}_i$. Mirroring the classic autoencoder, the VAE has two parts as well - the encoder, or the inference network, specifies a variational distribution over the latent space $q_\phi(\boldsymbol{r}|\boldsymbol{x})$ that approximates the true posterior distribution $p_\theta(\boldsymbol{r}|\boldsymbol{x})$ of the latent variables ($\boldsymbol{r}$) based on the observed data ($\boldsymbol{x}$). The decoder, or the generative network, models the generative process $x \sim p_\theta(\boldsymbol{x}|\boldsymbol{r})$ where $p_\theta(\boldsymbol{x}|\boldsymbol{r})$ is the probability distribution of the observed data ($\boldsymbol{x}$) given the latent variables ($\boldsymbol{r}$). Thus, the $\boldsymbol{r}$ acts as a sufficient statistic for the data: given sufficiently rich encoders and decoders, $\boldsymbol{r}$ captures all of the information of the original data $\boldsymbol{x}$.

To learn representations of the data, we thus need to learn the amortized (i.e., shared across all observations) encoder and decoder parameters, ϕ and θ, respectively. We do so by minimizing the Kullback-Leibler (KL) divergence between the true posterior and the approximate posterior, as a function of θ and ϕ. This minimization corresponds to maximizing the evidence lower bound, or

ELBO, given by:

$$ELBO(\theta, \phi) = \mathbb{E}_{\boldsymbol{r} \sim q_\phi(\boldsymbol{r}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{r})] - D_{KL}[q_\phi(\boldsymbol{r}|\boldsymbol{x})||p(z))] \tag{11}$$

Here, $D_{KL}[\cdot||\cdot]$ is the KL divergence. The first term of the above equation can be interpreted as maximizing reconstruction accuracy. The second term serves as a regularizer that penalizes estimates that are far from the prior. In other words, the log-likelihood encourages the decoder to reconstruct the data well, while the regularizer ensures that the encoder learns smooth latent representations of the input data. In the standard VAE model, the latent variables are not shared across data points, so the ELBO can easily decomposed into a sum where each term depends only on a single data point:

$$ELBO_i(\theta, \phi) = \mathbb{E}_{\boldsymbol{r}_i \sim q_\phi(\boldsymbol{r}_i|\boldsymbol{x}_i)}[\log p_\theta(\boldsymbol{x}_i|\boldsymbol{r}_i)] - D_{KL}[q_\phi(\boldsymbol{r}_i|\boldsymbol{x}_i)||p(z_i))] \tag{12}$$

This objective function can then be maximized using standard numerical methods.

F.2 Our Multimodal Autoencoder

We extend the previous VAE framework to accommodate the different modalities of TikTok posts. First, we use modality-specific encoders that capture the specifics of a given modality (e.g., audio). Next, we combine the modality-specific encoders using a neural network, which allows them to interact in a potentially nonlinear manner. This joint econder is then used to estimate the latent representation of a post. The decoder mirrors this structure. Given the highly structured nature of the modality-specific encoders and decoders, we refer to our proposed model as "very structured" VAE, or VS-VAE. Figure 4 in the main body shows the overall modeling framework, visually. We describe each component in more detail.

Encoder Recall the goal of the encoder is to infer the approximate posterior distribution of the latent representation $\mathbf{z}_i$ for video i, given its content, $\mathbf{x}_i$. In line with the typical assumption employed in a standard VAE, we assume a Gaussian approximation to the posterior, such that

$$p(\mathbf{z}_i \mid x_i) \approx q(\mathbf{z}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i I). \tag{13}$$

Thus, the encoder maps the data $\mathbf{x}_i$ to the variational parameters $\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i$. The encoder has two parts: first, each modality is encoded separately through a modality-specific encoder. We index modalities by $m = 1, \ldots, M$, and denote this first-stage modality specific encoding as:

$$\mathbf{h}_i^m = \text{ModalityEncoder}_{\phi_m}^m(\mathbf{x}_{im}), \tag{14}$$

where $\mathbf{h}_i^m$ is the intermediate representation learned from video i's modality m. Typically, these modality-specific encoders take the form of a deep neural network, parameterized by ϕ_m, whose architecture varies depending on the modality. Subsequently, the intermediate representations are concatenated together using a joint encoder, denoted by,

$$(\boldsymbol{\mu}_i, \log \boldsymbol{\sigma}_i) = \text{JointEncoder}_{\phi_0}([\mathbf{h}_i^1, \ldots, \mathbf{h}_i^M]), \tag{15}$$

where the function is again modeled using a deep neural network, parameterized by ϕ_0.

For each modality, we structure its encoder to reflect the unique structure of the data of that modality. We outline each of these architectures here:

- **Text:** The input to the text encoder is the sequential list of word embeddings from each of the three sources of text (video description, sticker text, and voice-over). To extract features from each of these word embeddings, we pass them separately through a Bi-LSTM layer, which is the typical choice to accommodate sequences of text, as it learns the meaning of a sentence by processing word sequences in both forward and backward directions (Schuster and Paliwal, 1997). The output of the Bi-LSTM is a vector of hidden states, which constitute the text-specific representation, $\mathbf{h}_i^{\text{Text}}$.

- **Image:** The input to the image encoder includes the vectors of VGG-19 features for each of the video frames, the 1000-D object vector, and the 2089-D visual sentiment vector. To extract features from the sequential VGG-19 features, we pass them through a LSTM layer that combines them into a single 4096-D VGG-19 vector while keeping the temporal relations between each frame. We then concatenate the hidden states of this LSTM with the other two input modalities, and pass it through several fully connected layers, to learn the video-specific representation, $\mathbf{h}_i^{\text{Video}}$.

- **Audio:** The input to the audio encoder is the 543-D audio feature vector, including the percentage measure of music universality, the music classes, and the ten standard acoustic features. We pass these through through multiple fully connected layers to get the audio-specific representation, $\mathbf{h}_i^{\text{Audio}}$.

- **Editing:** The input to the editing encoder is the 157-D editing feature vector, which we pass through a single fully connected layer to get the editing-specific representation, $\mathbf{h}_i^{\text{Editing}}$.

- **Joint:** Finally, the joint encoder takes each of the modality-specific representations, $\mathbf{h}_i^{\text{Text}}$, $\mathbf{h}_i^{\text{Video}}$, $\mathbf{h}_i^{\text{Audio}}$, and $\mathbf{h}_i^{\text{Editing}}$, and feeds them through a single-layered fully connected network to provide the variational parameters $\boldsymbol{\mu}_i$ and $\boldsymbol{\sigma}_i$.

Decoder The goal of the decoder is to reconstruct the observed data from the multimodal representation, $\mathbf{z}_i$. This reconstruction involves specifying a generative model for the data $\mathbf{x}_i$, parameterized as a function of the representation $\mathbf{z}_i$. We assume that, conditional on $\mathbf{z}_i$, each of the modalities is independent, and thus can be specified using a modality-specific decoder. Mirroring our notation for the encoder, we generically write,

$$\hat{\mathbf{x}}_i^m = \text{Decoder}_{\theta_m}^m(\mathbf{z}_i),$$

where $\hat{\mathbf{x}}_i^m$ indicates the parameters governing the data generating process for $\mathbf{x}_i^m$. For instance, for words, $\hat{\mathbf{x}}_i^m$ is the probability of seeing that word. For real-valued features, $\hat{\mathbf{x}}_i^m$ is the expectation of that value. The specific architectures for each modality's decoder are:

- **Text:** To generate text, the latent representation is first passed through a fully connected layer to create separate Bi-LSTM inputs for the three sources of text: video description, sticker text and voice-over. Then for each of them, we pass the Bi-LSTM outputs through a time distributed fully connected layer with softmax activation to get the probability of each word at each given time step.[20] Therefore, for each video, the output of the text decoder is a matrix of probabilities for every word and each position in the text. We employ categorical cross-entropy loss for the text reconstruction, which corresponds to modeling the text with

[20]The time distributed layer applies the same activation function to each of the time stamps.

a categorical likelihood.

- **Image:** The image decoder reconstructs the VGG-19 feature matrix, the object and the visual sentiment vector. To do so, we first pass $\mathbf{z}_i$ through multiple fully connected layers to obtain (1) the inputs for an LSTM, which is then used to reconstruct the VGG-19 features; and (2) the reconstructed object and visual sentiment vector means. We employ a mean squared error loss for the image reconstruction, which corresponds to a Gaussian likelihood for the image features.

- **Audio and Editing:** Both of these modalities use a single-layered fully connected network as decoders, together with the mean squared error loss for feature reconstruction, again corresponding to a Gaussian likelihood over the features.

Overall Loss To optimize the parameters θ and ϕ, we use the same ELBO loss as defined for the VAE in Equation 12, but extend it to capture the reconstruction losses for the multiple domains. Specifically, the loss for our VS-VAE model is as follows, where for simplicity, we suppress the notation of domain-specific parameters θ and ϕ in the right hand side parentheses:

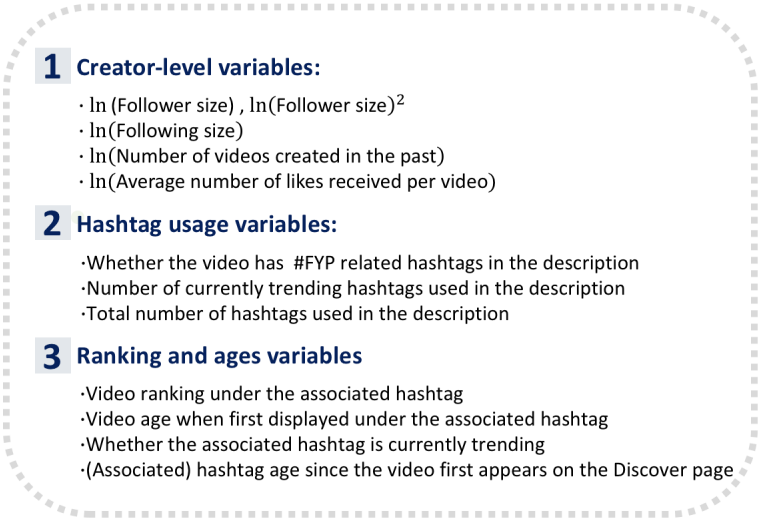
$$\mathcal{L}_{\text{VS-VAE}}(\theta, \phi) = \sum_{m} \mathcal{L}_{\text{recons},m} = \mathcal{L}_{\text{recons}_t} + \mathcal{L}_{\text{recons}_i} + \mathcal{L}_{\text{recons}_a} + \mathcal{L}_{\text{recons}_e} + \mathcal{L}_{KL}(\theta, \phi). \tag{16}$$

To minimize this loss, we use the Adam optimizer (Kingma and Ba, 2014). The model is trained on batch size 32 for 300 epochs with early stopping and a learning rate of 1e-5. To prevent overfitting, we use an L2-regularizer on the weights of our model. The specific details of the model architecture are given in Web Appendix B.

Web Appendix G: Predicting Video Popularity

To predict the popularity of a TikTok video, we proceed in two steps: the first part of our analysis aims to quantify general impression growth patterns, using only the non-content factors listed below. We refer to this step as "Stage 1" of our modeling framework. The results (described in more detail below) show that content unrelated features explain only a small part of the variance in the growth of impressions. We then leverage the content of videos as expressed by our VS-VAE representations to assess. This step is what we refer to as "Stage 2" of our predictive model. The good performance of our "Stage 2" prediction indicates that our representations can well capture the quality differences across the video content that drives its surprisingly good or bad popularity growth.

Figure 13: Non-content Factors



G.1 Stage 1: Baseline Model

We use a logarithmic growth specification as our baseline model, with the content-unrelated features as covariates. Intuitively, videos tend to have an initial period of fast impression growth, followed by relatively slow growth, which mirrors the assumptions of logarithmic growth. We estimate a Bayesian, multi-level logarithmic growth model, with post- and hashtag-level parameters, which are themselves predicted by the set of content-unrelated features. Let y_{ijt} denote the impression count for video j under hashtag i observed on day t, where $t = 1$ corresponds to the

video's first appearance under the hashtag.[21] We specify y_{ijt} as follows:

$$y_{ijt} = A_{ij} \log(t) + B_{ij} + \varepsilon_{ijt} \tag{17}$$

$$A_{ij} = \exp(a_{ij}),\ B_{ij} = \exp(b_{ij}) \tag{18}$$

$$\varepsilon_{ijt} \sim \mathcal{N}(0, \tau^2) \tag{19}$$

where a_{ij} is the log growth rate of video j under hashtag i and b_{ij} is its log starting impression. To capture the variation in growth rates and starting impressions, we assume a hierarchical, multi-level structure on a_{ij} and b_{ij}, such that:

$$a_{ij} \sim \mathcal{N}(\alpha_{0i} + [\boldsymbol{x}^u_{ij}, y_{ij1}]' \boldsymbol{\alpha}_i, \sigma^2_{\alpha,i}), \tag{20}$$

$$b_{ij} \sim \mathcal{N}(\beta_{0i} + \boldsymbol{x}^{u\prime}_{ij} \boldsymbol{\beta}_i, \sigma^2_{\beta,i}), \tag{21}$$

where $\boldsymbol{x}^u_{ij}$ is the set of non-content covariates for video j under hashtag i. In short, our multi-level formulation assumes that each video's growth parameters arise from a combination of hashtag-specific effects $(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)$ and the content-unrelated features for that video. The realized initial impression y_{ij1} is also included as an independent variable in the model of a_{ij}. Finally, each of these hashtag-level parameters is drawn independently from Gaussian priors:

$$\boldsymbol{\alpha}_i \sim \mathcal{N}(\boldsymbol{\mu}_\alpha, \sigma^2_\alpha I), \tag{22}$$

$$\boldsymbol{\beta}_i \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \sigma^2_\beta I), \tag{23}$$

Since we don't have any prior knowledge, we specify diffuse distributions for the set of hyperparameters $\tau^2, \sigma^2_{\alpha,i}, \sigma^2_{\beta,i}, \mu_{\alpha_0}, \mu_{\beta_0}, \sigma^2_{\alpha,0}, \sigma^2_{\beta,0}, \boldsymbol{\mu}_\alpha, \boldsymbol{\mu}_\beta, \boldsymbol{\sigma}^2_\alpha, \boldsymbol{\sigma}^2_\beta$. More specifically, we assume that the parameters for the intercepts $(\mu_{\alpha_0}, \mu_{\beta_0})$ and covariates $(\boldsymbol{\mu}_{\boldsymbol{\alpha}}, \boldsymbol{\mu}_\beta)$ are mutually independent. We assume a diffuse normal prior for both μ_{α_0} and μ_{β_0} and each element of $\boldsymbol{\mu}_\alpha$ and $\boldsymbol{\mu}_\beta$. In case of the variance terms, including both the estimation variance terms $\tau^2, \sigma^2_{\alpha,i}$ and $\sigma^2_{\beta,i}$ and the coefficient variance terms $\sigma^2_{\alpha,i}, \sigma^2_{\beta,i}, \sigma^2_{\alpha,0}, \sigma^2_{\beta,0}, \boldsymbol{\sigma}^2_\alpha$, and $\boldsymbol{\sigma}^2_\beta$, we follow the convention of specifying a gamma prior for the inverse of the variance of each parameter.

[21] If a video has been listed under multiple hashtags, we only consider its earliest appearance to avoid duplication.

Crucially, using Bayesian R^2 (Gelman et al., 2019), we find that non-content covariates are able to explain only about 33% of the variation in a_{ij} and b_{ij} across all hashtags. That nearly two-thirds of the variation is unexplained suggests that a large number of videos may exceed, or fall short of, their "destiny" — the performance we would expect by looking at who posted the video, and under what hashtag. In the following subsection, we use the *content* of the videos for demand prediction.

G.2 Stage 2: Residual Prediction based on the Modeled Video Content

We build neural network predictors to predict both the direction and value of the unexplained impression growth from Stage 1 based on the learned video representations from our VS-VAE framework.

To measure how much the realized popularity growth of a video is different from its estimated baseline in the previous subsection, we calculate its growth residuals defined as follows

$$a_{ij} - \hat{a}_{ij} \ , \ b_{ij} - \hat{b}_{ij} \tag{24}$$

where the parameter a_{ij} (b_{ij}) represents the logarithmic fitted growth rate (starting impression) we ultimately try to predict. The estimate $\hat{a}_{ij}$ (or $\hat{b}_{ij}$) represents the estimated value determined by only the non-content baseline factors from the first-stage estimation. The difference then captures a video's growth rate (or starting impression) that cannot be explained prior to accounting for its content. A positive difference refers to over-performance, meaning that its high-quality content helps it to realize a growth rate (or starting impression) higher than expected (hidden gem). A negative difference, on the other hand, refers to under-performance as a result of the video's lower-than-expected popularity growth (lemon) under its unsatisfactory content.

Mathematically, the difference takes the following form:

$$a_{ij} - (\mu_{\alpha_0} + \boldsymbol{X}_{ij}\boldsymbol{\mu}_{\alpha}) \tag{25}$$

$$b_{ij} - (\mu_{\beta_0} + \boldsymbol{X}_{ij}\boldsymbol{\mu}_{\beta}) \tag{26}$$

It is calculated based on the common parameters ($\alpha_0, \boldsymbol{\mu}_{\alpha}, \beta_0$ and $\boldsymbol{\mu}_{\beta}$) representing all hashtags in

our dataset, thus capturing not only the video-level but also the recurring content attributes resulting from belonging to a specific hashtag that help a video achieve its observed growth.

We use neural networks to flexibly capture the link between the video content and the unexplained impression growth. To classify over- and under-performing videos and predict the magnitude of their unexpected popularity (or lack thereof), we build sign classifiers and numeric predictors that take the learned video representations as input and provide the binary signs and the numeric values of the growth residuals, respectively, as an output. They consist of three fully connected layers with size 128, 64 and 1. We adopt the Rectified Linear Unit (ReLU) as the nonlinear activation function for the first two layers and softmax activation for the last layer of the sign classifiers and linear activation for that of the numeric predictors. They are trained by optimizing the binary cross-entropy and mean squared error loss, respectively.