Predicting Sales Lift of Influencer-generated Short Video Advertisements: A Ladder Attention-based Multimodal Time Series Forecasting Framework

Zhe Fu
The University of North Carolina at Charlotte
zfu2@charlotte.edu

Kanlun Wang
The University of North Carolina at Charlotte
kwang17@charlotte.edu

Jianfei Wang
Hefei University of Technology
wangjianfei9@126.com

Yunqin Zhu
University of Science and Technology of China
hasined@mail.ustc.edu.cn

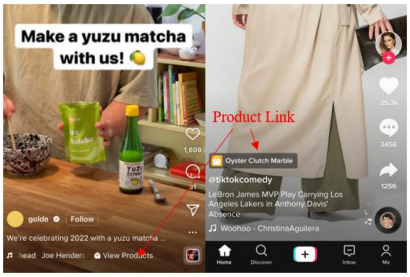
Abstract

With the growing popularity of video-sharing platforms and video influencers, influencer-generated short video advertisements (ISAs) have rapidly emerged as a crucial marketing tool. However, effectively predicting the sales lift of multiple ISAs presents significant challenges due to the multimodal content of ISAs and their impact of joint complexity on product sales. In this research, we design a novel time series forecasting framework that leverages ladder attention-based multimodal to predict the sales lift of multiple ISAs. Our framework, enriched by a novel ladder attention model and a customized LSTM-based time series forecasting model, addresses the challenges of predicting the sales lift of multiple ISAs. We conduct experiments using our proposed framework on a comprehensive dataset of ISAs collected from TikTok, and our results demonstrate superior performance in comparison to the baseline methods. This study not only offers a novel predictive tool in short video advertisement optimization but also serves as a guide in multimodal prediction in information systems and marketing research.

Keywords: Short video advertisements, Influencer marketing, Multimodal analysis, Attention mechanism, Time series forecasting

1. Introduction

The innovation of mobile technology (e.g., 5G and high-definition cameras) and social media have sparked the rise of short video platforms, such as Instagram Reels, YouTube Shorts, and TikTok. For instance, TikTok enables users to create, share, and discover short videos and attract over 150 million daily active users (Omar & Dequan, 2020). Among them, some have established significant accomplishments by creating and disseminating high-quality entertainment content (e.g., daily life, experiences, and opinions) on social media, which are recognized as influencers (Han et al., 2022). By incorporating product advertisement(s) into influencers' videos, products can expeditiously reach a large amount of foot traffic based on the extensive follower base of influencers and even proliferate to millions of online users through the platform's recommendation system. Influencer-generated Short video Advertisements (ISAs), which are short videos with marketing content produced by influencers and published on short video platforms, have proven to be highly effective in driving brand engagement and product sales (Leung et al., 2022). For example, users can shop directly from a shoppable in-feed video by tapping the product link or basket icon on Instagram and TikTok (See Figure 1).



(a) Instagram (b) TikTok

Figure 1. ISA examples from Instagram and TikTok

The primary objective of ISAs is to maximize the return on advertising investment. This return is manifested in the incremental sales conversion attributed to the impact of ISAs on e-commerce platforms, namely the *sales lift* of ISAs. Short video platforms have initially developed tracking tools (e.g., YouTube Analytics and TikTok Pixel) to help businesses track and monitor the conversion rate of marketing activities. However, tracking tools are incapable of predicting precise ISAs' sales lift, which in turn causes numerous business-related issues, such as relevant influencer selection, video content

URI: https://hdl.handle.net/10125/106726
978-0-9981331-7-1

management, determination of timing and scope for video marketing, and inventory management (Mallipeddi et al., 2022). Accurate forecasting of sales increases for ISAs addresses these challenges, offering valuable decision-making assistance in the burgeoning influencer marketing industry. As confirmed in 2022, the global influencer advertising market surpassed $16 billion in size and exhibited robust ongoing growth (Wies et al., 2022).

Furthermore, predicting the sales lift of multiple ISAs that are created for an individual product is challenging due to the highly heterogeneous multimodal content of videos and the joint complexity of multiple videos. First, businesses typically give influencers significant autonomy to create multimodal content (such as text, audio, and visuals) according to their preferences and personal branding, enabling the content to resonate effectively with their followers. (Leung et al., 2022). Thus, ISAs' multimodal content is highly heterogeneous due to discrepant influencer styles in content production, production resources, and photography skills (Y. Guo et al., 2022). The fusion of different modalities in a video can instigate disparate effects on product sales and poses a significant challenge in leveraging ISAs' multimodal content to predict their sales lift. Second, a short video platform commonly utilizes a collection of videos created by multiple influencers to collaboratively promote a product. Multiple ISAs for a given product are sequentially released and collectively contribute to influencing product sales. In addition, video content, release time, target audience, and other relevant dimensions of ISAs can also jointly pose an impact on sales life prediction.

Prior studies on ISA or live streaming have focused on the investigation of video features that affect consumer responses, such as multimodal content, influencer followers, and product quality (L. Chen et al., 2023; Liu et al., 2023; Fei et al., 2021; Koh & Cui, 2022). A limited stream of studies has attempted to predict the viewing duration or product click of a single ISA (R. Chen, 2022; Y. Guo et al., 2022), or prediction effect (Xi et al., 2023; Lin et al., 2023) and video popularity prediction (Ou et al., 2022) of live streaming. However, the research predicting the sales lift from multiple ISAs remains significantly underexplored. In addition, existing studies employed multimodal fusion techniques, such as non-neural network-based models (Shan et al., 2020), gate attention-based multimodal (Du et al., 2022; G. Wang et al., 2023), and concatenation-based multimodal (Jaafar & Lachiri, 2023; Yang et al., 2023), to capture the interactions between various modalities, resulting in enhanced accuracy for downstream predictions. Nevertheless, in the context of short videos, these approaches are unsuitable as they do not properly account for the intrinsic correlations between products and videos, especially in terms of the collective impact of multiple videos.

To solve the above challenges, we propose the following research questions: 1) How do we leverage multi-modality information in improving the performance of multiple ISAs' sales lift prediction, compared to the state-of-the-art models? and 2) How does the exclusion of modality deteriorate or improve the performance of multiple ISAs' sales lift prediction?

Moreover, we develop a Ladder Attention-based Multimodal time series forecasting framework (LAMM) for multiple ISAs' sales lift prediction. Drawing upon multimodal learning, attention mechanism, and LSTM architecture, LAMM automatically extracts text-, audio-, frame-, and motion-level embeddings from ISAs, learns their interactions using a novel ladder attention model, and dynamically predicts the sales lift of multiple ISAs in an end-to-end framework. To evaluate the LAMM, we collected a large-scale ISAs Dataset (ISAD) from TikTok, consisting of 77,708 ISAs created by 33,877 influencers, across 1,809 products. The objective of this study is to predict a product's sales for the forthcoming day by utilizing product-related videos, as well as influencer and product metadata.

This research makes three-fold contributions as follows. First, we propose a novel ladder attention-based multimodal time series forecasting framework to predict the sales lift of multiple ISAs. By incorporating a proposed ladder attention model and a customized LSTM-based time series forecasting model, the framework significantly outperformed the baseline methods. Second, this study identifies the essential characteristics within multiple modalities linked to short-video advertisement sales. This forms a foundational benchmark for future research in the fields of multimodal analysis and short video advertising. Third, this study makes valuable contributions to the field of information systems and marketing research by providing a publicly available dataset for multimodal data analytics. Specifically, the dataset is constructed by sales-relevant, entertainment-driven, and influencer-generated short videos.

2. Related work

Despite its exponential expansion, the impact of ISAs on sales lift has received limited attention. To bridge this gap, we investigate studies within a broader research domain that delve into influencing factors and predictive approaches for assessing the effects of online video advertising, with a particular focus on the pivotal attributes of video in the context of short video advertising.

2.1. Influencing factors of online video advertising effects

Online video advertising has demonstrated a positive influence on consumer purchasing behavior, particularly on traditional websites and e-commerce platforms (Kumar & Tan, 2015; Addo et al., 2022). Some studies further highlighted the substantial impact of videos' multimodal content on consumer reactions, such as Song et al. (2021) revealed that the complexity of visuals and colors in e-commerce advertising videos affects product click behaviors, and Y. Guo et al. (2022) found that product description, product demonstration, content pleasure, and content aesthetics are four key determinants of viewing durations.

Similar to e-commerce short video advertisements, ISAs on social media, especially those on short video platforms have a positive influence on consumer responses, such as product sales (Ge et al., 2021), purchase intention (Filieri et al., 2023), advertising clicks (R. Chen, 2022), and advertising engagement (L. Chen et al., 2023). The influential factors of ISAs can be divided into four levels based on their respective contributions to video advertising, including video, creator, product, and user (Xiao et al., 2023). Among them, creator-, product-, and user-level factors typically involve numeric or text-based information, such as follower count and comment, yet video-level factors are complex due to their multiple modalities (i.e., visual, audio, text, numeric) and are the most important and essential components of short video platforms (R. Chen, 2022). Given the complexity of video-level factors, short video platforms face a cold-start problem that refers to the lack of user interaction information and product history information (R. Chen, 2022). Multimodal content in videos can reduce the role of cognitive efforts and positively reshape product-related quality signals, thereby facilitating persuasion (Chang et al., 2023) and reducing product uncertainty (Dimoka et al., 2012), ultimately promoting product sales (Addo et al., 2022). More importantly, it has been evidenced that the interaction between different modalities has an impact on advertising effectiveness. For example, Pozharliev et al. (2022) suggested that consumers use both text information (i.e., the description of a product) and image information (i.e., a picture of an influencer using a product) to jointly evaluate the effectiveness of advertisements.

In a nutshell, the above studies emphasize the significant impact of the multimodal content of short videos, which motivates utilizing the multimodal content of ISAs to predict sales lift. Nonetheless, these investigations fail to thoroughly explore the multimodal aspects of video content, including elements like frames, motion, audio, and textual information. The impact of this abundant yet diverse information on online video advertising outcomes remains unclear.

2.2. Predicting online video advertising effects

Predictions of online advertising effects, including metrics like click-through rate and user engagement, have consistently garnered attention in both research and practical applications (Gharibshah & Zhu, 2022). An accurate prediction can enable businesses to optimize their marketing strategies effectively, thereby enhancing their competitiveness in the digital marketplace. Despite the numerous studies that have been conducted on predicting online advertising effects, existing research has overlooked the impact of image and video features which are distinct characteristics of multimedia advertising on their predictive models. The incorporation of multimodal features from videos holds a wealth of valuable information that not only improves the prediction of online video advertising effects but also assists in mitigating the cold-start problem associated with new video advertisements. (Ou et al., 2022). To this end, R. Chen (2022) proposed a multimodal cooperative learning framework to predict product clicks by automatically extracting multimodal features and capturing the relations between different modalities; Y. Guo et al. (2022) proposed a multimodal multitask framework to predict video advertisements' viewing response by distilling shared, special, and conflicted information from multimodal features; and Ou et al. (2022) proposed a multimodal and temporal attention fusion framework to predict user engagement by representing and combining multimodal features. These methods extracted multimodal features of videos through pre-trained models (e.g., ResNet, VGGish, and BERT), and designed multimodal fusion methods to improve prediction performance. However, their primary emphasis was on precursor variables of product sales, including user engagement, view durations, and product clicks, without considering the actual product sales, which are the ultimate goal of advertising efforts. This lack of attention to real sales data means that these models may inadvertently overlook significant trends and patterns, ultimately resulting in suboptimal performance (Addo et al., 2022).

The nature of social media platforms, with their focus on entertainment and social engagement, gives rise to notable distinctions between short video advertisements on social media and e-commerce platforms. These differences encompass various aspects, including the creators involved (i.e., influencers versus brands or sellers), the topics addressed (i.e., entertainment-oriented versus product-oriented), the level of standardization (i.e., weak versus strong), and the joint impact of multiple factors (i.e., strong versus

weak), among others. More specifically, businesses commonly engage multiple influencers and provide them with creative autonomy to craft multimodal content (including text, audio, and visual elements) that aligns with the unique personal positioning of each influencer. This approach facilitates the creation of content that resonates with followers and enhances the influencer's credibility (Leung et al., 2022). In contrast to the curated and limited number of video advertisements created by businesses, the collective impact of multiple ISAs is inherently more intricate and diverse. This complexity arises from factors such as varying quality, content, and a larger quantity of ISAs. Therefore, predicting the sales lift of ISAs becomes a significant challenge due to these distinctive characteristics.

To guide the design of more effective multimodal methods, we further review short video multimodal analysis studies and representative multimodal fusion methods. Short video analysis studies involving two typical types of tasks, including short video understanding such as classification (Li et al., 2023), tagging (X. Wang et al., 2022), highlight detection (Z. Guo et al., 2022), as well as short video applications such as video recommendation (Almeida et al., 2022), popularity prediction and video engagement (Lu et al., 2023). Although short video analysis involves various tasks, it is usually treated as a multimodal fusion problem that contains visual, audio, and text modalities. In addition to short videos, the emergence of multimodal data in various fields is driving the development of multimodal fusion methods, which include but are not limited to non-neural network-based models (Shan et al., 2020), gate attention-based multimodal (Du et al., 2022; G. Wang et al., 2023), and concatenation-based multimodal (Jaafar & Lachiri, 2023; Yang et al., 2023).

In summary, existing online advertising effect prediction methods have not addressed the joint complexity of multiple videos affecting product sales, making it difficult to accurately predict the product sales of multiple ISAs.

3. Methodology

3.1. Problem formulation

We formulate the sales lift prediction problem as a regression task that predicts the sales lift values for each product. Let v denote a set of videos related to a product, which contains multiple modalities (i.e., text, audio, motion, and frame) and v serves as the source of input information for building regression models. The task can be formulated as learning a regression model for sales lift prediction at time t (see Equation (1)):

$$y_t = f(\Theta, V_{t-1}) \quad (1)$$

where y_t denotes a predicting sale of a target product g at time t, Θ denotes the set of parameters of the regression model $f(\cdot)$, and $V_{t-1} = \{v_1, v_2, \dots, v_{t-1}\}$ is the set of videos related to the target product g before time $t-1$. More importantly, the proposed framework integrates the information of both multimodal contents and multiple publishers. The overall framework of our proposed method is illustrated in Figure 2, which consists of three components: multimodal feature extraction, multimodal feature fusion, and sales lift prediction.

3.2. Multimodal feature extraction

We pre-processed the ISAD dataset and extracted numerical features of products, influencers, and videos (see Table 1). To encompass a wider range of indicators that signify the connection between ISAs and product sales, we conducted additional analysis by extracting visual, audio, and text modality features from short videos. These features were obtained using established pre-trained models, aligning with the methodologies outlined in the pertinent literature (e.g., R. Chen, 2022; Y. Guo et al., 2022; Ou et al., 2022).

We focused on extracting visual features from two distinct perspectives: motion features, which are associated with the movement of objects, and frame features, which capture the static characteristics within each frame. To extract motion features, we utilized a methodology inspired by the work of Y. Guo et al., (2022), where multiple frames per 3.2 seconds were inputted into the TimeSformer model (Bertasius et al., 2021) simultaneously. This process allowed us to extract dynamic embeddings with a dimension size of 128. To align the number of dynamic motion, frame, and audio features for each video, we utilize MViTv2 (Li et al., 2022) to learn frame features and VGGish (Hershey et al., 2017) to learn acoustic deep features with a 128-dimension size of every 3.2 seconds. Furthermore, we extracted dynamic motion, frame, and audio features from all videos to capture a comprehensive set of features. Specifically, we focused on the first 20 segments of each video, corresponding to a duration of 64 seconds. This approach was chosen considering that the dataset consisted of 97% of videos with a duration of less than 64 seconds. For videos with a length shorter than 20 segments, we pad the vectors of the remaining segments with zeros. Finally, the BERT embedding technique (Cui et al., 2021) is adopted to embed each word in video titles into 128-dimensional vectors.

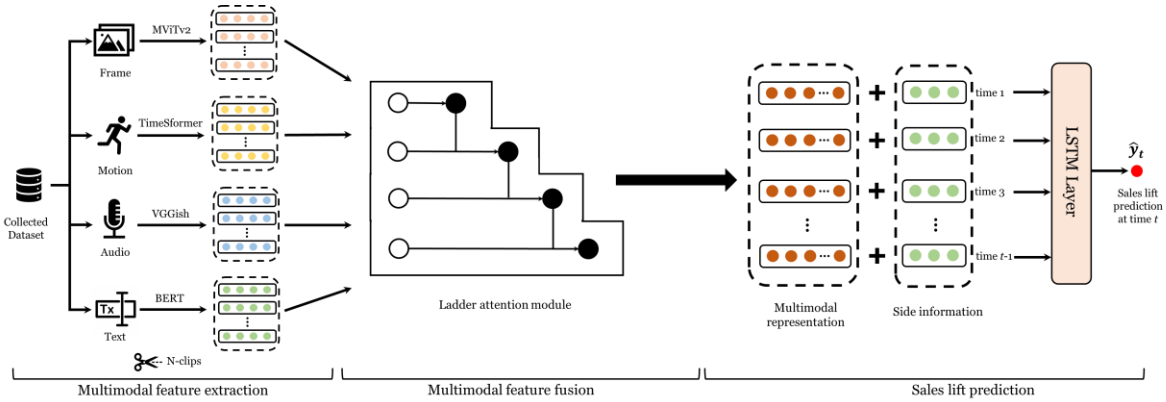


Figure 2. A ladder attention-based multimodal time series forecasting framework (LAMM)

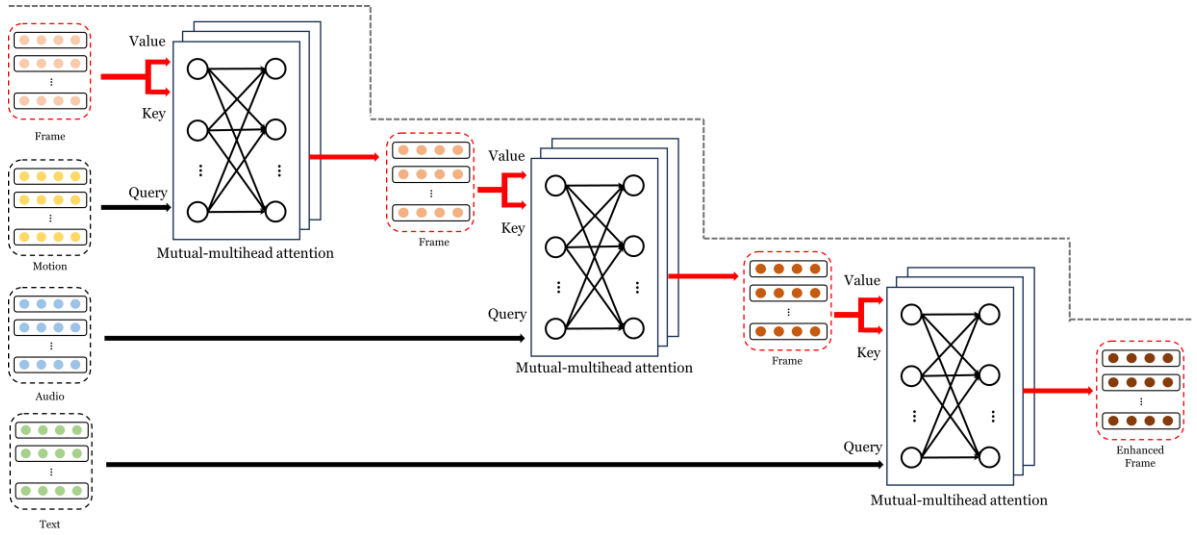


Figure 3. Ladder attention model for multimodal feature fusion

3.3. Multimodal feature fusion

Among the four modalities in the videos, the frame modality is the main body of the videos, which contains rich clues of ISAs in predicting sales of the products. The frame modality can be complemented and emphasized by the text, audio, and motion. For example, the title and the audio commentary of a video can help the users to better understand the main topic of the video. Therefore, it is beneficial to focus on the text, audio, and motion modalities as auxiliary clues and complementary information for product sales estimation. As a result, we adopt the attention mechanism to integrate the text, audio, and motion modality into frame modality by enhancing the representation of each frame segment. As shown in Figure 3, we designed a ladder attention model to aggregate the text, audio, and motion information into the frame modality asynchronously. The basic component of the proposed ladder attention model is the mutual-multi-head attention layer. Given a target Modality A and an auxiliary Modality B, we leverage the representation of Modality B at each segment as the Query (Q_B) of the attention layer and the representation of Modality A at each segment as Key (K_A) and Value (V_A). The enhanced representation of Modality A $h_{A\leftarrow B}$ is calculated by Equation (2):

$$h_{A\leftarrow B} = softmax(\frac{Q_B K_A^T}{\sqrt{d}})V_A \qquad (2)$$

where d is a 128-dimension representation of modality. In Equation 2, by leveraging the information provided by Modality B, we can calculate the weight for each segment of Modality A. This allows us to generate an enhanced hidden representation of Modality A, incorporating the insights derived from Modality B. Furthermore, we additionally designate the frame as the target modality, and in each iteration, we fuse the text, audio, and motion information with the frame representation to enhance its overall quality

and richness of information. Finally, a new sequence of hidden representations of the frame segments is generated from the multimodal feature fusion component. To represent the video, we further aggregate the frame segment vectors through an average pooling layer to generate an overall representation of h_v (see Equation (3)).

$$h_v = Ave([h_1^{f\leftarrow t,a,m}, h_2^{f\leftarrow t,a,m}, \ldots, h_n^{f\leftarrow t,a,m}]) \quad (3)$$

where $h_i^{f\leftarrow t,a,m} \in R^{1\times d}$ denotes frame representation of the i^{th} segment enhanced by text, audio, and motion.

3.4. Sales lift prediction

In the multimodal feature fusion component, we have obtained a final multimodal representation vector h_v for each video related to the product. However, during our analysis, we discovered that each product encompasses multiple pertinent videos, and the information associated with these videos, such as daily like count, daily share count, and daily comment count, exhibits dynamic changes over time. To address the above two challenges, we first concatenate the aforementioned side information associated with the videos m_{vt} with the video representation h_v to generate a time-dependent video representation h_{vt} (see Equation (4))

$$h_{vt} = Concat([h_v, m_{vt}]) \quad (4)$$

We model our sales lift prediction task as a time-series prediction problem by considering multiple active videos for a product on a daily basis and aggregating all the active video representations related to the product by a mean operation to generate a general representation of all the videos related to the product at time *t* (see Equation (5)).

$$h_t = Ave([h_{v1}, h_{v2}, \ldots, h_{vk}]) \quad (5)$$

where *k* is the total number of active videos at time *t*. Therefore, the final prediction on the product sales can be calculated based on the general multimodal information sequence $H_{t-1} = [h_1, h_2, \ldots, h_{t-1}]$ before time *t*. In this paper, we leverage the Long-Short Term Memory (LSTM) network to process the general multimodal sequence H_t and make predictions on product sales (see Equation (6)).

$$\hat{y}_t = LSTM(H_{t-1}) \quad (6)$$

where $\hat{y}_t$ is the predicted sales value at time *t*. We chose the Mean Squared Error (MSE) as the loss function (see Equation (7)):

$$L = \frac{1}{|T|}\sum_{t\in T}(\hat{y}_t - y_t)^2 + \mu||\Delta|| \quad (7)$$

where T denotes the days for sales lift prediction, and $\mu||\Delta||$ is parameter-specific regulation hyper-parameters to prevent overfitting. By minimizing the loss value calculated in Equation (7), we optimize the model to generate prediction results.

4. Experiments

4.1. Data collection

We collected a large-scale ISAs dataset from TikTok, a leading global short video platform (Omar & Dequan, 2020). TikTok influencers have up to 100 million followers and earn up to $5 million a year. Our focus is primarily on Douyin, the Chinese counterpart of TikTok, owing to its well-established e-commerce ecosystem centered around influencer video advertising. Within this ecosystem, Douyin Shopping functions as an integrated e-commerce feature within short video advertisements. It enables users to make direct product purchases from merchants on the Douyin e-commerce platform while simultaneously watching the short video advertisement. Meanwhile, Xingtu, as an influencer marketplace within the ecosystem, enables brands and product sellers to contract with influencers for short video advertising. By 2022, Xingtu had attracted over millions of influencers, brands, and product sellers.

To begin, we initially identified the popular products listed on Xingtu's hot list on March 2, 2023. Subsequently, we proceeded to collect data related to videos, products, and influencers for these identified products on a daily basis over a two-week period, spanning from March 2, 2023, to March 15, 2023. Our choice of a two-week period is rooted in the following considerations. Firstly, we observed that the daily increments in video collections, comments, likes, and shares exhibit a declining trend, ultimately stabilizing during the second week. This pattern suggests that there exists a two-week window of influence for sales generated by new videos. Secondly, with the daily influx of product-associated videos steadily increasing, both the volume of data and the challenges associated with data collection are proportionally escalating. To mitigate the potential impact of outliers in product sales and video counts, we have opted to exclude products that fall within the top 1% in terms of sales and video numbers. Finally, we obtained a cleaned dataset, namely ISAD, which comprises a total of 77,708 ISAs generated by 33,877 influencers. This dataset encompasses 1,809 products across 13 different categories.

Table 1 presents the descriptive statistics for variables at the product, influencer, and video levels. The "daily video count" for a product denotes the

quantity of new ISAs linked to that specific product on a particular day. The findings reveal an average daily addition of three new ISAs per product, reaching a peak of 52. This highlights the prevalence of multiple ISAs associated with the same product on short video platforms, emphasizing the difficulty in addressing the intricately combined influence of multiple ISAs on product sales.

Table 1. Descriptive statistics for product-, influencer-, and video-level variables

Levels	Variables	Mean	Min	Max
Product	daily sale count	58.9	0	995
	daily price (¥)	59.7	2.9	3499.0
	daily rating	89.1	71.40	100.00
	daily video count	3.0	0	52
	history sale count	4132.7	16	225658
	history video count	149.6	2	1080
	type (category)	13		
	age	29.7	1	99
	daily follower count	69609.6	65	25453260
Influencer	daily like count	697037.2	0	1686056000
	city (category)	First level (2885), New first level (4310), Second level (6937), Third level (7916), Four level (5565), Five level (2885), Others (4900)		
Video	daily collect count	7.8	0	9417
	daily comment count	5.9	0	1963
	daily like count	45.1	0	33393
	daily share count	6.0	0	8925
	create hour	11.5	0	23
	length	17.1	2.5	865
	height (category)	1920p (62921), 1280p (5953), 1080p (2726), Others (6108)		
	width (category)	1080p (64027), 720p (6290), 1920p (2062), Others (5329)		
	type(category)	14		

Note: Product types include Toys and Games (212) category, Sports and Outdoor Equipment (207), Home and Textiles (180), Shoes and Bags (147), Kitchen and Home Appliances (141), Clothing and Underwear (168), Fresh Produce (169), Beauty and Skincare (129), Maternity and Baby Products (105), Daily Necessities (104), 3C Electronics (82) and Jewelry and Accessories (81). Video types include Home life (158769), Food (92679), Fashion (146818), Parent-child (114166), Fitness (13429), Random clips (11319), Technology (10808), Sports (7502), Automotive (7214), Humanities and social sciences (5855), Agriculture and rural areas (5614), Entertainment (9761), Music and dance (687), Medical health (2185).

4.2. Baseline models

To evaluate the proposed LAMM model for sale prediction, we selected three baseline models, including concatenation-, gate attention-, and traditional machine learning-based methods. For the concatenation-based methods, we remove the proposed ladder attention model and fuse the multimodal features by simply concatenating the representations of all four modalities (i.e., text, audio, motion, and frame). For the gate attention-based methods, we further add a self-attention layer to the top of the concatenated multimodal representations for adjustment. In addition, we leverage LSTM and Multilayer Perceptron (MLP) layers for sales lift prediction respectively for both concatenation-based methods and gate attention-based methods. For the traditional machine learning-based methods, we selected the Logistic Regression (LR) model.

4.3. Evaluation metrics

We adopted a set of widely used evaluation metrics to measure the predicted results, including Mean Average Error (MAE) and Mean Squared Error (MSE). Moreover, we adopt five-fold cross-validation to evaluate the model performance, with an 80/20 partition for training and testing respectively.

4.4. Experiment settings

The collected dataset was utilized to train both the proposed LAMM framework and the baseline models. During training, we employed a learning rate of 10^{-3}, a hidden dimension of 128 for all modalities, a dropout rate of 0.2, and a regularizer decay of 10^{-3} across all models. The Adam optimizer was utilized for training all the models. As for model-specific hyperparameters, we set the number of heads in the mutual-multi-head attention layer to 3.

5. Results

5.1. Model performance

To assess the effectiveness of our proposed model, we conducted a comparative analysis of the performance between the LAMM model and the baseline models. The performance metrics of the models and the results of paired sample t-tests are presented in Tables 2 and 3, respectively. The tables display the performance values for each fold, with the best results in each column highlighted in bold.

The results show that our proposed LAMM model significantly decreased the MAE in comparison to the Gate_LSTM (t=-3.393; df=4; *p*=.027), Gate_MLP (t=-3.380; df=4; *p*=.019), Concat_LSTM (t=-4.482; df=4; *p*=.011), Concat_MLP model (t=-2.914; df=4; *p*=.044), and LR models (t=-6.427; df=4; *p*=.003). In addition, the model performances evaluated by MSE are similar to that of MAE. Our proposed LAMM model significantly decreased the MSE compare with Gate_LSTM (t=-3.042; df=4; *p*=.038), Gate_MLP (t=-3.619; df=4; *p*=.022), Concat_MLP (t=-3.183; df=4; *p*=.033), and LR models (t=-20.421; df=4; *p*=.000), yet yields a marginal effect in comparison to Concat_LSTM model (t=-2.752; df=4; *p*=.051).

Table 2. Model performance comparison on MAE between our proposed model and baseline models

Models	Fold					Mean	Sig.
	K1	K2	K3	K4	K5		
LAMM	**0.602**	**0.575**	**0.631**	**0.567**	**0.644**	**0.604**	-
Gate_LSTM	0.603	0.587	0.641	0.584	0.651	0.613	0.027
Gate_MLP	0.606	0.598	0.654	0.608	0.667	0.627	0.019
Concat_LSTM	0.607	0.593	0.648	0.590	0.653	0.618	0.011
Concat_MLP	0.603	0.604	0.648	0.607	0.656	0.624	0.044
LR	1.056	1.258	1.294	0.989	0.916	1.103	0.003

Table 3. Model performance comparison on MSE between our proposed model and baseline models

Models	Fold					Mean	Sig.
	K1	K2	K3	K4	K5		
LAMM	**0.976**	**0.918**	**1.044**	**0.905**	**1.027**	**0.976**	-
Gate_LSTM	1.020	0.929	1.075	0.950	1.028	1.000	0.038
Gate_MLP	1.013	0.924	1.069	0.950	1.034	1.000	0.022
Concat_LSTM	1.015	0.920	1.068	0.950	1.034	0.997	0.051
Concat_MLP	1.007	0.927	1.066	0.952	1.036	0.998	0.033
LR	2.121	1.913	2.012	1.931	1.863	1.968	0.000

Table 4. The ablation study of LAMM

Model	MAE	MSE
LAMM	**0.604**	**0.976**
w/o frame	0.628	1.043 ↑
w/o text	0.625	1.039 ↑
w/o audio	0.611	0.988
w/o motion	0.619	1.009

Notes: The best results in each column are bolded, and ↑ indicates a model error increase of more than 5% relative to the original LAMM model.

5.2. Ablation analysis

The results are averaged by five-fold outputs shown in Table 4 and show that incorporating all four modality information in terms of frame, text, audio, and motion achieves the best performances across all the metrics, including MAE and MSE. After ablating frame, the model increased by 3.97% in MAE (t=-2.862; df =4; p=.023), 6.86% in MSE (t=-4.150; df =4; p=.007); after ablating text, the model increased by 3.48% in MAE (t=-2.255; df =4; p=.044) and 6.45% in MSE (t=-2.557; df =4; p=.031); after ablating audio, the model increased by 1.16% in MAE (t=-3.31; df =4; p=.015) and 1.23% in MSE (t=-2.041; df =4; p=.055); after ablating motion, the model increased by 2.48% in MAE (t=-2.320; df =4; p=.041) and 3.38% in MSE (t=-2.417; df =4; p=.036).

6. Discussions

The prevalence of short videos as a primary medium for online entertainment and information has resulted in the rise of video influencers as influential figures in diverse domains of interest. The impact of ISAs on product sales is an area of research that is still evolving, despite their growing significance in marketing.

In this research, we propose a novel ladder attention-based multimodal time series forecasting framework for predicting the product sales lift of multiple ISAs. The superior performance of our framework was validated by a large-scale empirical ISVD dataset. The proposed framework offers marketers and advertisers a valuable tool to optimize their short video advertising strategies and maximize their return on advertising investment. With accurate predictions provided by the framework, influencers can also leverage this information to create short video content that holds greater advertising value. Moreover, our extensive ablation experiment carries substantial promise in evaluating the influence of integrating various modalities into the model development process. This proves especially beneficial in amplifying the effectiveness of short video advertisement sales promotion and has wider implications for time-series multi-modality forecasting. It is evident that video frames emerge as the most crucial element, with text following closely behind. This phenomenon can be elucidated by the audience's inclination to focus on short videos, as both casual online users and subscribers tend to pay more attention to video introductions or beginnings.

This study paves the way for further research. First, one important factor that may influence product sales is the recommendation mechanisms employed by short video platforms. However, it is worth noting that our study collected data over a relatively short time period and focused on a limited number of products without considering the potential impact of recommendation effects. This limitation should be countered in future research. Second, the proposed framework utilizes a mean operation to aggregate all the video representations associated with the product, thereby overlooking the variations among the videos, which could potentially impact the accuracy of predictions. It is recommended to explore more advanced time series forecasting methods that can capture the temporal dynamics of the video data.

References

Addo, P. C., Akpatsa, S. K., Nukpe, P., Ohemeng, A. A., & Kulbo, N. B. (2022). Digital analytics approach to understanding short video advertising in digital marketing. *Journal of Marketing Theory and Practice*, *30*(3), 405-420.

Almeida, A., de Villiers, J. P., De Freitas, A., & Velayudan, M. (2022). The complementarity of a diverse range of deep learning features extracted from video content for video recommendation. *Expert Systems with Applications*, *192*, 116335.

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? *Proceedings of the 38th International Conference on Machine Learning*, 813-824.

Chang, H. H., Mukherjee, A., & Chattopadhyay, A. (2023). More voices persuade the attentional benefits of voice numerosity. *Journal of Marketing Research*, 002224372211341.

Chen, L., Yan, Y., & Smith, A. N. (2023). What drives digital engagement with sponsored videos? An investigation of video influencers' authenticity management strategies. *Journal of the Academy of Marketing Science*, *51*(1), 198-221.

Chen, R. (2022). Multimodal cooperative learning for micro-video advertising click prediction. *Internet Research*, *32*(2), 477-495.

Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3504-3514.

Dimoka, Hong, & Pavlou. (2012). On product uncertainty in online markets: Theory and evidence. *MIS Quarterly*, *36*(2), 395.

Du, Y., Liu, Y., Peng, Z., & Jin, X. (2022). Gated attention fusion network for multimodal sentiment classification. *Knowledge-Based Systems*, *240*, 108107.

Fei, M., Tan, H., Peng, X., Wang, Q., & Wang, L. (2021). Promoting or attenuating? An eye-tracking study on the role of social cues in e-commerce livestreaming. *Decision Support Systems*, *142*, 113466.

Filieri, R., Acikgoz, F., & Du, H. (2023). Electronic word-of-mouth from video bloggers: The role of content quality and source homophily across hedonic and utilitarian products. *Journal of Business Research*, *160*, 113774.

Ge, J., Sui, Y., Zhou, X., & Li, G. (2021). Effect of short video ads on sales through social media: The role of advertisement content generators. *International Journal of Advertising*, *40*(6), 870-896.

Gharibshah, Z., & Zhu, X. (2022). User response prediction in online advertising. *ACM Computing Surveys*, *54*(3), 1-43.

Guo, Y., Ban, C., Liu, X., Goh, K. Y., Peng, X., Yang, J., & Li, X. (2022). Short-video marketing in e-commerce: Analyzing and predicting consumer response. *ICIS*.

Guo, Z., Zhao, Z., Jin, W., Wang, D., Liu, R., & Yu, J. (2022). TaoHighlight: Commodity-aware multi-modal video highlight detection in E-Commerce. *IEEE Transactions on Multimedia*, *24*, 2606-2616.

Han, X., Wang, L., & Fan, W. (2022). Cost-effective social media influencer marketing. *INFORMS Journal on Computing*, ijoc.2022.1246.

Hershey, S., Chaudhuri, S., Ellis, D. P. W., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, et al. (2017). CNN architectures for large-scale audio classification. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 131-135.

Jaafar, N., & Lachiri, Z. (2023). Multimodal fusion methods with deep neural networks and meta-information for aggression detection in surveillance. *Expert Systems with Applications*, *211*, 118523.

Koh, B., & Cui, F. (2022). An exploration of the relation between the visual attributes of thumbnails and the view-through of videos: The case of branded video content. *Decision Support Systems*, *160*, 113820.

Kumar, A., & Tan, Y. (Ricky). (2015). The demand effects of joint product advertising in online videos. *Management Science*, *61*(8), 1921–1937.

Leung, F. F., Gu, F. F., & Palmatier, R. W. (2022). Online influencer marketing. *Journal of the Academy of Marketing Science*, *50*(2), 226–251.

Li, Y., Liu, S., Wang, X., & Jing, P. (2023). Self-supervised deep partial adversarial network for micro-video multimodal classification. *Information Sciences*, *630*, 356–369.

Li, Y., Wu, C.-Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., & Feichtenhofer, C. (2022). MViTv2: Improved multiscale vision transformers for classification and detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4794-4804.

Lin, Q., Jia, N., Chen, L., Zhong, S., Yang, Y., & Gao, T. (2023). A two-stage prediction model based on behavior mining in livestream e-commerce. *Decision Support Systems*, 114013.

Liu, L., Fang, J., Yang, L., Han, L., Hossin, Md. A., & Wen, C. (2023). The power of talk: Exploring the effects of streamers' linguistic styles on sales performance in B2B livestreaming commerce. *Information Processing & Management*, *60*(3), 103259.

Lu, S., Yu, M., & Wang, H. (2023). What matters for short videos' user engagement: A multiblock model with variable screening. *Expert Systems with Applications*, *218*, 119542.

Mallipeddi, R. R., Kumar, S., Sriskandarajah, C., & Zhu, Y. (2022). A framework for analyzing influencer marketing in social networks: Selection and scheduling of influencers. *Management Science*, *68*(1), 75-104.

Omar, B., & Dequan, W. (2020). Watch, Share or Create: The Influence of Personality Traits and User Motivation on TikTok Mobile Video Usage. *International Association of Online Engineering*. 121-137.

Ou, N., Yu, L., Li, H., Du, Q., Xiang, J., & Gong, W. (2022). MTAF: Shopping guide micro-videos popularity prediction using multimodal and temporal attention fusion approach. *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4543-4547.

Pozharliev, R., Rossi, D., & De Angelis, M. (2022). A picture says more than a thousand words: Using consumer neuroscience to study instagram users' responses to influencer advertising. *Psychology & Marketing*, *39*(7), 1336-1349.

Shan, G., Lina, Z., & Dongsong, Z. (2020). What reveals about depression level? The role of multimodal features at the level of interview questions. *Information & Management*, *57*(7), 103349.

Song, D., Wang, S., Ou, C., Chen, X., Liu, R., & Tang, H. (2021). How do video features matter in visual advertising? An elaboration likelihood model perspective. *ICIS*.

Wang, G., Ma, J., & Chen, G. (2023). Attentive statement fraud detection: Distinguishing multimodal financial data with fine-grained attention. *Decision Support Systems*, *167*, 113913.

Wang, X., Gan, T., Wei, Y., Wu, J., Meng, D., & Nie, L. (2022). Micro-video tagging via jointly modeling social influence and tag relation. *Proceedings of the 30th ACM International Conference on Multimedia*, 4478-4486.

Wies, S., Bleier, A., & Edeling, A. (2023). Finding goldilocks influencers: How follower count drives social media engagement. *Journal of Marketing*, 87(3), 383-405.

Xi, D., Tang, L., Chen, R., & Xu, W. (2023). A multimodal time-series method for gifting prediction in live streaming platforms. *Information Processing & Management*, *60*(3), 103254.

Xiao, L., Li, X., & Zhang, Y. (2023). Exploring the factors influencing consumer engagement behavior regarding short-form video advertising: A big data perspective. *Journal of Retailing and Consumer Services*, *70*, 103170.

Yang, Y., Qin, Y., Fan, Y., & Zhang, Z. (2023). Unlocking the power of voice for financial risk prediction: A theory-driven deep learning design approach. *MIS Quarterly*, 47(1), 63-96.